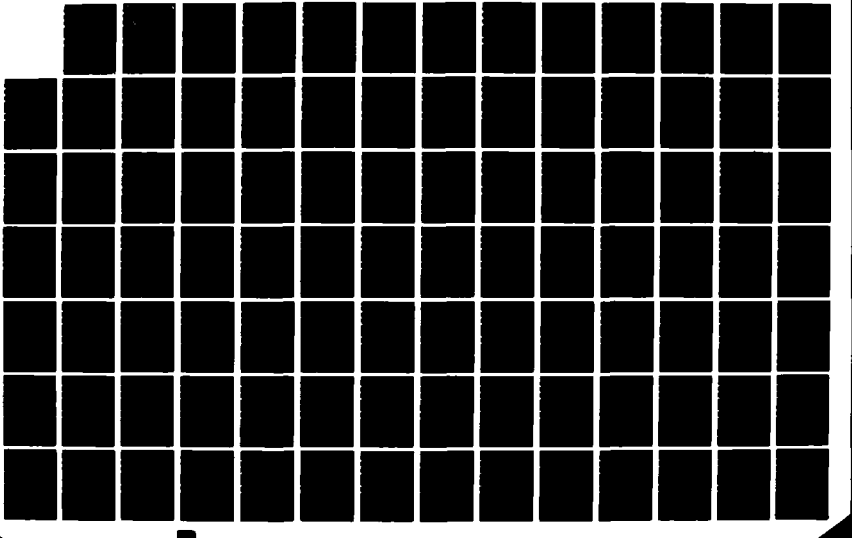
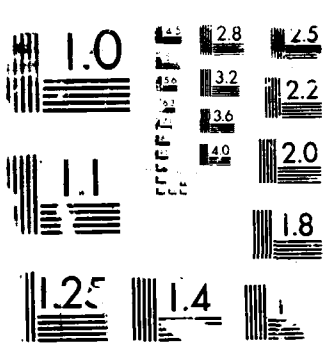
AD-A189 451 COMPUTER AIDED FAST TURNAROUND LABORATORY FOR RESEARCH IN VLSI (VERY LARG (U) STANFORD UNIV CA CENTER FOR INTEGRATED SYSTEMS J D MEINDL ET AL 31 MAY 87 1/3

UNCLASSIFIED MDA903-84-K-0062 F/G 9/1 NL





Computer Aided Fast Turnaround Laboratory for Research in VLSI

Final Report

on

DARPA Contract No. MDA903-84-K-0062

by

The Stanford University

Center for Integrated Systems

October 1, 1983 to May 31, 1987

Principal Investigator:
Professor James D. Meindl
CIS 105
Stanford University
Stanford, California
94305
(415)725-3622

Project Leader:
Dr. John Shott
CIS 211
Stanford University
Stanford, California
94305
(415)725-3715

Contents

Chapter 1

Introduction

The Stanford Center for Integrated Systems (CIS) is a leading national effort to forge closer links between industry and academia. It is a genuine university-government-industry research onsortium involving nineteen major U.S. corporations. They have contributed $14.25 million for construction of a new CIS building. Its principal features are VLSI design and fabrication laboratories including 10,000 square feet of vibration-free clean rooms for a computer aided/Automated fast turn-around laboratory (CAFTAL). Due to the novel blending in the CIS of eminent academic and industrial scientists from both the VLSI design and fabrication disciplines, the CAFTAL is uniquely positioned effectively to address the future technological needs of fast turn-around fabrication for VLSI.

The principal objectives of the CAFTAL for VLSI are: 1) application of cutting edge computer science and software systems engineering to fast turn-around fabrication in order to develop more productive and flexible new approaches; 2) fast turn-around fabrication of optimized VLSI systems achieved through synergistic integration of system research and device research in aggressive applications such as superfast computers, and 3) investigation of physical limits on submicron VLSI in order to define and explore the most promising technologies.

To make a state-of-the-art integrated circuit process more manufacturable, we must be able to understand both the numerous individual process technologies used to fabricate the complete device as well as the important device, circuit and system limitations in sufficient detail to monitor and control the overall fabrication sequence. Specifically, we must understand the sensitivity of device, circuit and system performance to each important step in the fabrication sequence. Moreover, we should be able to predict the manufacturability of an integrated circuit before we actually manufacture it.

The salient objective of this program is to enable accurate simulation and control of computer-integrated manufacturing of ultra large scale integrated (ULSI) systems, including millions of submicron transistors in a single silicon chip, through creative application of computer science and software engineering. Accurate simulation of a complete manufacturing line for ULSI systems demands a tightly integrated set of research tasks to produce the necessary software. Computer control of a manufacturing laboratory that fabricates working prototypes of ULSI systems will verify the simulation programs and convincingly demonstrate the essential unity of all research tasks.

To pursue its focused objective effectively, this program is organized in five cross-disciplinary projects.

Knowledge Tools for The Manufacturing Automation project is designing new high-level knowledge-based programming languages, FABLE, using several levels of abstraction to obtain

the complete specifications for a sequence of semiconductor manufacturing operations, and it will demonstrate that these specifications are correct by applying them to control both automated manufacturing and simulation. With the FABLE language we will be able to describe processes so that they are understandable, designable, debuggable, portable and runnable by a suitable computer-based automation system. There are five tasks within the Knowledge Tools project. Two tasks are concerned with techniques for representing knowledge about manufacturing. The goal of the first task, Process Specification for Automated Semiconductor Fabrication, is to implement a version of the specialized representation language Fable for specifying semiconductor manufacturing processes. To complement this first task, we will be working with an industrial collaborator on a second task called 'Knowledge Representation in Processes'. This project concentrates on generic manufacturing knowledge. The project will gather a manufacturing vocabulary, develop a knowledge representation to express the manufacturing vocabulary, and develop user interfaces to permit manufacturing experts to encode their knowledge. The third task, Process Entry and User Interfaces, will provide a set of graphics-based tools to support research in knowledge-based factory automation. The fourth task, Process Automation System Architecture, will concentrate on the hardware and software infrastructure of the automated factory. This task will build on the SECS communication protocols to provide a robust distributed computer environment for managing and controlling the automated fabrication line. The last of the five Knowledge Tools tasks is Intelligent Interfaces to Semiconductor Processing Equipment. The goal of this task is to create a 'processing workstation', incorporating a processing equipment and a highly capable engineering workstation with AI capabilities which will provide a user-friendly control panel for the processing equipment, as well as facilities for editing, simulating, debugging, and running recipes.

The Factory Modeling and Management project is researching semiconductor factory-level productivity relationships including how throughput and turnaround time depend on operating policy, on the nature of demand, and on the gross characteristics of equipment and supervisory control mechanisms. Investigators from the Stanford Graduate School of Business, the Industrial Engineering and Engineering Management Department of the School of Engineering, and industry are addressing these issues. This project is supported by a parallel SRC program on Manufacturing Science. The work is included in this report because of its relevance to the CAFTAL goals and activities.

The Simulation project is incorporating equipment, process, device and circuit models in a far-reaching effort to develop a manufacturing line simulator that inputs a FABLE specification for producing a complex semiconductor chip and outputs a set of histograms describing key parametric distributions for the chip assuming a specific fabrication line and process sequence.

In the Technology project we are developing a new modeling discipline-semiconductor manufacturing equipment modeling. Perhaps the greatest obstacle to simulating a semiconductor manufacturing line is an almost total lack of physical models to describe process-parameter variations resulting from characteristics (e.g. the geometry) of particular machine designs. Existing models used in process simulators such as SUPREM are highly generic and assume that a specific local environment is replicated across an entire wafer, from wafer-to-wafer in a batch and from batch-to-batch. This assumption is an invaluable simplification in developing process models per se but neglects *the* fundamental manufacturing problem of parameter distributions and their causes. Equipment modeling will address this problem. The Patterning project is concerned with modeling optical, electron beam and other lithography equipment. The Etching project is focusing on modeling plasma, reactive ion and other etching equipment. The Deposition and Redistribution

project includes modeling the equipment for physical and chemical vapor deposition, oxidation and diffusion, rapid thermal processing, and ion implantation. The general approach to each project begins with the development of new test structures and measurement tools to improve understanding of manufacturing parameter distributions. This experimental data is then be used as a guide to the formulation of predictive physical models for various machines. Agreement between measured and theoretical distributions will verify the models and serve as a basis for defining in concert with manufacturers new machine concepts for future generations of equipment.

The Diagnostics and Yield Modeling project is addressing a critical requirement for the advancement of semiconductor manufacturing science and technology-accurate interpretation of the results of manufacturing steps, whether singly or in concert. This entails using innovative test chips and methodologies such as process deconvolution structures, arranged in logic tree fashion, that will enable yield diagnosis by partitioning the universal set of possible offenders into increasingly smaller subsets. The accomplishments in the CAFTAL program over the past three years are summarized below.

This report describes the work in these areas during the course of this research program.

1.1 Fabrication Automation

1.1.1 System Implementation

- The CIS building has been completely wired for two separate ethernet systems. One for office, CAD, and Measurement lab use and one for laboratory use only.

- The entire laboratory clean room space has been wired for SECS-I communication to equipment. This involves at least 8 modular plug connections for each laboratory wall. This connections are compatible with the complete SECS-I (Semiconductor Equipment Control Specification), including the optional power supply lines for optical isolation.

- An LSI-11/23 computer running a stripped down set of UNIX 2.9 code is now running as an interface between the ethernet and the measurement system IEEE-488 bus. This allows the development within a well controlled operating system of software for rapidly transferring large amounts of measurement data from the HP-4062 measurement system to any computer system on the building or campus network for analysis.

- A MicroVax-I system running 4.3 UNIX now has a copy of the Berkeley MicroLab system for laboratory monitoring and control. This machine is running the SUN Network File Server code which allows transparent access and transfer of data between computer systems. This has required modification to the Berkeley system, opens up a number of possibilities of improvement and will serve as an initial framework for our monitoring and control work.

- A MicroVax-I running 4.2 UNIX and using a CAMAC (Computer Automated Measurement and Control) system was installed and wired to several parts of a sputtering system. After developing and debugging the CAMAC interface to the MicroVax this allows control and monitoring of several parts of the machine. Because this computer is on the laboratory ethernet access to the machine is also possible remotely and even via telephone hookups to other machines.

1.1.2 Process Specification

The overall objective of this research was to design and implement Fable, a language for describing semiconductor processes. When completed, Fable will be used as a target language for process design and as a executable language for process implementation and control. At present, a Fable design has been completed and prototype implementations of Fable exist.

Fable supports OBJECTS, for abstract representation of entities such as equipment and materials. The Fable object system supports *multiple views*, so that different aspects of the process information may be emphasized for different purposes. Fable provides a procedural language for specifying the steps in a process. Fable provides a non-procedural language, based on the Grid formalism [5], to efficiently encode the web of relationships between objects and processes.

The specific achievements of the process specification project over the last three years are:

- The Grid sub-system was implemented in Modula-2, as described by Ossher in his Ph.D. thesis.
- A prototype of the Fable object system, with multiple views, was implemented this year in Common Lisp.
- A prototype of the procedural part of Fable, emphasizing the need for intelligent exception handling, is currently under development in Common Lisp.
- 1 Ph.D. Student Graduated.
- 1 Journal Article.
- 3 Conference Presentations.

The work done in this project has received significant attention in the industry. Texas Instruments has contributed personnel and a TI Explorer Lisp machine (worth approximately $70,000). Schlumberger's collaborative project is closely aligned with Fable and is highly complementary; Schlumberger will make their CIMANTIC/Object Flow system available for use by the Fable project. IntelliCorp has contributed a copy of the KEE knowledge engineering tool.

1.2 Process Simulation

Stanford has a long history of work in the area of process and device simulation. Much of this work has been performed under the leadership of Profs. R. W. Dutton and J. D. Plummer under DARPA sponsorship. Significant accomplishments in the areas of process and device simulation include:

- Development of the SUPREM [23] family of process simulators. The current release of this family of simulators has been shipped to over 400 remote industrial, government, and university sites world wide.
- Publication of over 300 journal articles based on the process physics which has gone into the development of an improved understanding of processing and process simulation.

- The granting of over 20 PhD degrees to students involved in the development of the process models and simulation tools which have gone into the SUPREM program.
- Notable areas in which the SUPREM research program has improved our understanding of process physics include:
  - Enhanced oxidation of heavily doped silicon regions.
  - The kinetics of thin (< 500 Angstrom) oxide layers on silicon.
  - The behaviour of dopants in polycrystalline silicon including their impact on resistivity and grain size.
  - The role of point defects (interstitials and vacancies) on the diffusion of impurities in silicon.
  - The role of oxidation and nitridation in generating point defects which subsequently affect the diffusion of underlying impurities.
  - The physical mechanisms responsible for the gettering of metallic impurities from silicon wafers.
  - The impact of process parameters on the resulting fixed charge densities at the Si/SiO_2 interface.
  - The properties of W and WSi_2 as potential gate electrode and contact metallurgies.
- Development and release of the PISCES-II [26] two-dimensional device simulator which includes solution of Poisson and the one- and two-carrier continuity equation for non-planar devices.

More recently, we have initiated a program to develop a manufacturing-level process and device simulator based on the SUPREM and PISCES simulation engines, respectively. In particular we:

- Completed a preliminary draft of the language to be used for automating device simulation using PISCES.
- Implemented portions of the device simulation grammar using *lex* and *yacc* to become familiar with the tools to be used in the development of of the device simulation compiler.
- Completed the draft language for high-level process simulation.
- Specified and implemented the first generation user interface and compiler for high-level process simulation.

1.3 New Processes, Devices, and Circuits

1.3.1 Lithography

The major accomplishments in microlithography to date under the current DARPA contract can be summaried as follows:

- Stable reticle and mask generation capability for various optical exposure tools has been established, using the Mebes I electron beam system. In addition to conventional e-beam resists, photoresists such as Shipley 2400 were used in the mask making process, which yielded superior resolution, CD control and low defect density demanded by advanced VLSI patterning applications. This facility has been providing key support for various advanced research projects in microelectronics at Stanford, e.g. the Ultra High Speed GaAs Device Technology effort led by Professor James Harris.

- The 1/8 μm contract with Perkin Elmer EBT Division has retrofitted a 1/64 μm stage interferometer and a magnetic column shield onto our Mebes I system to achieve better placement accuracy. A write scan switch to improve write scan linearity at smaller address sizes is being engineered. The system is currently being evaluated for improved line edge roughness. This contract is scheduled to end in July 1986.

- Using the Mebes system and the Ultratech 900 stepper, lithography processes have been successfully integrated into the 2 μm CMOS technology developed in this laboratory for multi-project circuit fabrication. The patterning and overlay performances are expected to be adequate for further scaled design rules down to 1.25 μm.

- A data link established between the Ultratech stepper and the central laboratory computer, in conjunction with associated software development, has facilitated more flexible stepper operation and simplified data management. Better understanding in the long-term stepper stage drift behavior and the precision of the alignment system was obtained from data collected through this interface.

- A trilayer resist process has been developed for electron beam direct write lithography using the Mebes. It has been applied to patterning contact holes of sizes down to 1/4 μm for a contact resistance study [50]. Experiments are being prepared to incorporate this process in submicron-gate MOS device fabrication using e-beam and optical lithographic modes in a mix-and-match manner.

- A vote-taking lithography technique [51] has been developed to eliminate the effect of mask defects in photolithography. Experimental evaluation, using a novel etching technique which isolates lithography-induced pattern defects, has demonstrated its effectiveness in yield improvement. This method is an attractive alternative to the tedious and costly effort of ensuring low mask defect density, and is especially useful in small volume fast turnaround situations such as ASIC fabrication. Ultratech Stepper has expressed a strong interest in it.

- A new strategy for mask/wafer pattern defect inspection has been developed. It detect defects through direct feature extraction using template patterns determined by the circuit design rules, and hence eliminates the need for any reference comparison. Content-addressable memories (CAMs) are used for its implementation, thereby facilitating real-time throughput. The CAMs for detecting both random defects and dimensional errors have been designed with our 2 μm CMOS design rules, and one of them is currently being fabricated in our laboratory.

- The apparatus for Langmuir-Blodgett film growth has been installed and a variety of films have been grown in controllable thicknesses. The exposures of brassidic acid films with scanning electron microscope and tunneling electron microscope demonstrated subtenth micron

resolution. Thin films of protoporphyrin IX dimethyl ester, which is a thermally stable organic semiconductor material, are being characterized for their electrical properties. In addition, cadmium arachidate films have been incorporated in MIS structures to increase the barrier heights.

- 3 published papers.

- Major equipment donations have been made to Stanford as a result of our unique position in lithographic research among academic institutions. These included a GCA 6300 stepper, an Ultratech 900 stepper (50%), a Nanometrics Cwikscan IIIE SEM, and a Silicon Valley Group resist processing system.

1.3.2 Interconnections and Contacts

The overall objective of this research was to investigate conducting and insulating materials, fabrication processes, and device structures for multilevel interconnections and contacts in sub-micron VLSI, so that advances in integrated electronics can continue. Specifically, we have investigated low pressure CVD of tungsten and tungsten silicide, alloys of aluminum with titanium and other refractory metals and polycrystalline silicon to obtain better device structures and to overcome the problems of VLSI outlined above. The specific achievements during the past three years of this program are:

- Low-pressure chemical vapor deposition of tungsten silicide has been done and the properties of the deposited films have been studied to determine the process compatibility and suitability to form gate electrodes and interconnections in MOS VLSI applications.

- A hot wall tubular reactor has been developed to do low pressure chemical vapor deposition of tungsten. Technology to do selective deposition of W on Si, metals and silicides has been developed. Thin films of W have been shown to provide reliable contacts to shallow junctions with low resistance and low leakage. W was found to be a good barrier against Si diffusion into Al. Schottky diodes have been fabricated by selective CVD of W on N type Si and the technology has been used to fabricate PMOS transistors for latch-up free CMOS. Selective CVD of W has also been investigated to provide a low resistivity shunt over the high resistivity shallow junctions.

- Layered structures and homogeneous alloy films synthesized by sputter deposition have been developed for use in a VLSI multilevel interconnect technology. We have demonstrated in this work that aluminum alloyed with silicon and titanium, or layered with titanium offers advantages over current technological materials for interconnections in integrated circuits. The impact of this research is that low resistivity, hillock free, dry etchable electromigration resistant metal films can be fabricated and used in VLSI multilevel interconnects.

- A new technology has been developed to fabricate high performance MOS transistors in fine grain polycrystalline silicon by hydrogenation of the film by ion-implantation. The devices have been shown to improve the performance of the circuits. Feasibility of the technology has been proven for SRAMs.

- Techniques to accurately extract specific contact resistivity (P_c) from contact resistance data (V/I) have been developed by 2-D simulations of the measurement structures. It has been shown that in past P_c had been overestimated because of the use of 1-D models which can't take into account parasitic resistance associated with contact resistance. Accurate values of P_c have been determined for PtSi, Pd_2 Si, Al and W contacts to Si.
- 3 Ph.D. Students Graduated.
- 11 Journal Articles.
- 12 Conference Presentations (4 Invited).
- 1 best paper award for the paper authored by D. Gardner, T. Michalka, K. Saraswat, J. McVittie, T. Barbee and J. Meindl. "Aluminum Alloys with Titanium, Tungsten and Copper for Multilevel Interconnections," presented at the 1st International IEEE VLSI Multilevel Interconnection Conference, New Orleans, June 1984.
- 1 Patent has been granted on Al/Ti/Si layered films.

The work done in this project has received tremendous attention in the industry. Through joint projects, visits, seminars and talks the results have been picked up by Texas Instruments, G.E., Motorola, Intel, IBM, AMD, Kodak, Rockwell, Harris, Gould/AMI, H.P., Fairchild, Xerox, IDT, INMOS, VLSI Technology and Phillips/Signetics. Currently these companies are in constant touch with us to implement our research in their manufacturing.

1.3.3 Etching

The major accomplishments in etching under the current DARPA contract can be summarized as follows:

- Etch processes for poly-Si, silicon nitride, oxide, aluminum alloys and oxide planarization have been established to support full device fabrication down to the 1.5 μm dimension level. This etch capability has been used to fabricate a wide variety of devices and test structures.
- A novel electrical end point scheme was developed for oxide contact etching where present optical methods fail at small total contact areas. This method which is more than 20X more sensitive than other methods relies on monitoring an injected current flowing from the etch plasma through the open contacts and out the backside of the wafer.
- An optical system was designed to measure of spatial variations in optical emission in the plasma above a wafer during etching. By using argon and an ion probe for calibration, the spatial distribution of reactive neutrals and effective electron density can be determined. A kinetic model for the generation and loss of reactive neutrals has been developed and agreeds well with measured concentration distributions.
- Ion surface damage is a major problem in anisotropy etching of silicon. To eliminate this damage an low ion energy anisotropic silicon etch process was developed. This process gives high etch rates, high selectivity to oxides, low undercut and none of the usual etch damage.

- A novel approach to studying the effect of ion energy on etching has been developed without affecting the rf discharge in a parallel plate type etcher. This method uses a local metal grid held above part of the wafer to suppression ion bombardment by effectively increasing the sheath thickness and thereby increasing the ion collisions in the sheath. This approach should be extremely valuable for studying the properties of inhibitor layers which are responsible for anisotropy in a number of etch processes.

- An example of using etching techniques to fabricate novel test structures is our study of two-dimensional oxidation where anisotropic silicon etching was used to create cylindrical rings with radii down to 0.5 μm. The oxidation of these rings allowed observation of the effect of curved surfaces on oxidation in a geometry for which a closed form analytical solutions could be obtained. For this work an oxidation model was obtained which truly takes into account the effect of stress at corners.

- 3 Ph.D. Students to Graduate in 1986.

- 2 Journal Articles.

- 7 Conference Presentations.

- 1 Best student paper award for the paper authored by D.B. Kao, J.P McVittjie, W.D. Nix, and K.C. Saraswat," Two-Dimensional Silicon Oxidation Experiments and Theory," presented at the International Electron Device Meeting, Washington, D.C., December, 1985.

- 1 Patent has been applied for on the Electrical End Point Method.

- Finally, we have established close ties with etching groups at many at the CIS and SRC member companies. The electrical end-point method has been set-up at both HP and Xerox, and we are transferring plasma diagnostic tools to both Intel and Fairchild. In the area of equipment vendors, cooperative work has resulted in etch system donations from Drytek, Applied Material and Branson.

1.3.4 Thin Dielectric Films

The down-scaling of metal-oxide-semiconductor devices motivated by the continuing increase in the integration density of integrated circuits requires a substantial reduction in oxide thickness in the field-effect transistor gate, DRAM storage capacitor, and nonvolatile-memory tunnel dielectrics. The technological and reliability problems associated with silicon dioxide in the very thin regime emphasize the need for alternative high-quality insulators and new growth techniques. Silicon nitride, nitrided oxide (nitroxide), and oxidized nitride (oxynitride) grown by thermal and plasma nitridations have been investigated and are proposed as the best available alternatives. New techniques developed to grow these films are thermal nitridation in a radio-frequency-heated reactor, rapid thermal oxidation and nitridation in a lamp-heated system, and low-temperature microwave nitrogen-plasma nitridation. Using these techniques the following accomplishments have been achieved:

- The kinetics of the thermal nitridation of silicon and silicon dioxide in an ammonia ambient have been analyzed. The electrical characteristics of metal-insulator-semiconductor devices

with thermal silicon-nitride and nitroxide gate dielectrics are examined, and the results indicate the excellent electrical stability of the silicon-nitride devices because of very low carrier trapping. The interface transition from nitride to silicon is abrupt, and the morphology and roughness of the interface are comparable to the oxide/silicon interfaces.

- We have demonstrated that rapid thermal nitridation is a possible approach to the use of higher temperatures for very short times. Application of this process to silicon dioxide creates nitrided barrier layers at the surface and interface, increases interfacial charge densities, and slows the generation rate of new surface states resulting from electrical stress. The formation kinetics of these nitrogen-rich layers have been correlated to the electrical behavior of the rapidly grown nitroxides.

- A new plasma-nitridation technique based on nitrogen plasma generated by microwave discharge reduces nitridation temperature and enhances the growth kinetics. Besides silicon we have demonstrated the feasibility of growing insulating films on germanium and GaAs.

- 1 Ph.D. student graduated.

- 6 Journal articles.

- 13 Conference papers (2 invited).

- 2 Patents have been filed.

- The work done on this project has received a lot of attention in the industry and the results have been picked up by several of them. In particular Xerox, Gould/AMI, Intel, G.E. and AT&T have implemented results of our research in their programs through contacts with us. Others may have used the reports and publications, however, we don't have the details of that.

1.4 Electrical Wafer Testing

The parametric testing and diagnostics task, initiated 1/85, has set its sights on the development of automated tools for the acquisition and interpretation of data needed for IC process characterization, control, and the frequently flawed and seemingly intractable process problem debugging. In pursuit of these objectives we have initiated a vigorous development effort in in-process monitors, end-of-process monitors, and data collection, reduction, and interpretation strategies.

The in-process monitor development effort, initially aimed at the critical processing areas of patterning and doping, has led to the development of a unique technique [40] that permits separation of defects generated by the photolithographic process from defects generated by the etching process, hitherto inseparable. We have also developed an ion implant dose monitor [39] and have initiated a collaborative effort with Varian to develop implant channeling [52], shadowing, and surface charging [53] monitors.

Last year's work in end-of-process monitors has been focused on developing an interrelated set of process decomposition test structures which permit an unambiguous determination of the locations and densities of random defects in a dual metal, n-well CMOS process, and a novel approach to IC yield prediction [38]. Our approach has been well received by the IC industry, as evidenced by

numerous requests for the test chip manual, invitations for seminars from GE, HP, AMI, Reticon, Schlumberger, and Harris, and collaborative projects with HP, Reticon, Schlumberger, and VLSI Standards. We are now focusing our efforts on the development of test structures for process decomposition of CMOS circuit simulation parameters, and on the development of a complete set of test structures for the determination of flexible, optimum design rules.

The above work has led to three publication [38,39,40] and five conference presentations, three of them invited.

To support the process monitor development efforts, we have obtained equipment and software donations from HP, Prometrix, Micromanipulator, Suss, BBN, and ENHANSYS, totaling over $400,000.00. These donations have allowed us to establish a first-rate measurement and characterization laboratory, which now includes the HP4062 Parametric Test System coupled to HP9000 computer running ENHANSYS statistical data analysis software, an EDAX equipped Hitachi S-570 SEM, an IBT scanning optical microscope, the Prometrix Omnimap system, and numerous probers and other analytical equipment.

The effort in automated data reduction and interpretation has now been initiated with the addition of two new Ph. D. level students, and will be bolstered in the coming months by two visiting industrial scholars from Intel and Siemens, complementing the existing team of two senior research associates, two senior Ph. D. students, and one technician.

Chapter 2

Fable: Knowledge Tools for Manufacturing Automation

Byron Davies, Jay M. Tenenbaum, Ernie Wood, Lia Adams, Paul Asente, Evan Kirshenbaum, Lanre Amos

The Fable project has made significant progress over the past year. We learned important lessons from our first efforts in this area, and are forging ahead with renewed energy and excitement.

Our recent accomplishments are as follows:

1. We have identified the important lessons learned from our initial designs and implementations of Fable.
2. We have broadened Fable to capture more of the knowledge necessary to describe and carry out semiconductor manufacturing.
3. We have identified and acquired a very powerful set of hardware and software tools to support the continued development of Fable.
4. We have identified and started four applications projects that will serve both to determine specific requirements for Fable and to implement specific parts of Fable to satisfy these requirements.
5. We are conducting, for the first time, a Stanford course in "Automation of Semiconductor Manufacturing" (CS412/EE391).
6. We have started a nationwide electronic mailing list, IC-CIM, for discussion of topics in computer-integrated manufacturing of integrated circuits.

2.1 Lessons

The principal lessons learned from Fable's first three years were the following:

1. The Fable problem is largely a problem in representing, acquiring, and using a broad variety and large amount of knowledge about semiconductor manufacturing. The procedural model initially assumed by Fable was not adequate to express the knowledge required. We are now using a more powerful and more general model – knowledge representation.

2. The Fable problem is more difficult than we anticipated. Its solution requires state-of-the-art tools in symbolic computing, knowledge representation and inference. We have acquired and are now using such tools, including AI workstations (TI Explorers) and knowledge engineering tools (KEE and SimKit).

3. The Fable problem requires more collaboration than we anticipated. Its solution requires close interaction between experts in IC processing, computer science, and manufacturing, and equally close interaction between university and industry. We are working hard to establish both kinds of collaboration.

2.2 Knowledge and Representation

Much of the knowledge required by an automated fabrication system cannot be effectively encoded as procedures. To reason about a fabrication process, for example, requires information not just *of* the process, but also information *about* the process. An example of knowledge that is difficult to express procedurally is the following:

> Recipe CM142 is very similar to Recipe CM101, which was begun 24 hours earlier. If the initial parametric data from CM101 are marginal or worse, it is recommended that CM142 be suspended until the problems are identified and corrected.

Although it is *possible* to encode such information procedurally, more declarative representations of such information can be easier to understand and modify.

The automated fabrication line needs to know about more than just the processes that run on it. To execute the processes, the fab line needs to have knowledge about equipment, materials, schedules, and preferences, for example. We will need to use a broad variety of knowledge representation mechanisms to effectively capture knowledge about such a broad variety of entities.

2.3 Knowledge Tools

To efficiently develop prototypes of systems to gather and use such knowledge for automation, we need powerful tools. Thanks to Texas Instruments and IntelliCorp, we now have access to a very powerful software development environment, incorporating the TI Explorer workstation and the KEE knowledge engineering environment. In one project, this environment has enabled us to do, in one month, more than could have been accomplished in six months in a more conventional Unix environment. We expect this productivity to extend to the entire Fable project.

2.4 Applications Projects

CIS researchers have identified five new projects related to automation of semiconductor manufacturing, and have begun to work on them.

The first such project involves semiconductor factory simulation. A group of three students have implemented a queueing model simulation of a semiconductor fabrication line, and are investigating algorithms for scheduling the simulated line. The students are developing a process specification

language (a prototype of the procedural component of Fable) to permit the simulated line to run realistic fabrication recipes.

The second project is focused on developing intelligent processing equipment. The general idea is to closely couple a piece of semiconductor manufacturing equipment, such as an ion implanter, with a state-of-the-art AI workstation. The linkage will give the workstation direct access to the sensors and controls of the processing equipment. The linkage permits and encourages the development of knowledge-based software to support:

1. interactive monitoring and control of the processing equipment through a graphical interface;
2. high-level communication between host computer and the augmented processing equipment;
3. automated diagnosis of equipment and process problems;
4. automated monitoring and control of the recipe running on the processing equipment; and
5. design and simulation of process steps to run on the equipment.

Part of this second project involves developing languages (as part of Fable) for describing recipes specific to particular classes of processing equipment, such as ion implanters.

A third project is investigating expert systems for automatically interpreting electrical measurements from test structures. When the electrical measurements deviate significantly from those predicted by circuit simulation (using SPICE, for example), we would like an expert system to tell us which physical parameter is likely to be responsible for the deviation, and the amount of the deviation .

The fourth task, Process Automation System Architecture, will concentrate on the hardware and software infrastructure of the automated factory. This project is building on the SECS communication protocols to provide a robust distributed computer environment for managing and controlling the automated fabrication line.

The last of the five Knowledge Tools projects is Intelligent Interfaces to Semiconductor Processing Equipment. The goal of this task is to create a 'processing workstation', incorporating a processing system, such as a sputtering system or an ion implanter, and a highly capable engineering workstation with AI capabilities. The AI workstation will provide a user-friendly control panel for the processing system, as well as facilities for edi ing, simulating, debugging, and running recipes. This system will, we hope, serve as a prototype for a future generation of more intelligent unit processing systems.

We believe that these knowledge tools will allow us to effectively address the very complex problems of automating manufacturing processes. Flexibility in factory automation requires that knowledge about the factory, the processes used, and the products produced be encoded in explicit representations to be used by a wide variety of reasoning and simulation programs. The tools described here will provide such representations as well as facilities to acquire and use the knowledge needed to automate manufacturing processes.

2.5 Collaboration

Automating semiconductor manufacturing will require expertise from a number of fields, including semiconductor processing, computer science, and IC manufacturing. The Fable project welcomes the participation of people from all these fields.

Automating semiconductor manufacturing will require collaboration between researchers and implementors in both university and industry. The Fable project welcomes the participation of industrial colleagues and encourages students to obtain first-hand experience with industrial semiconductor fabrication lines.

2.6 New Automation Course

"Automation of Semiconductor Manufacturing" (CS412/EE391) is being offered for the first time this fall at Stanford. The course includes lectures on semiconductor manufacturing, existing automation in the semiconductor industry, object-oriented databases, scheduling, knowledge representation and expert systems for semiconductor manufacturing, automated interpretation of test structures, and other AI topics. To learn from our industrial colleagues and to help them learn about topics in semiconductor automation, we have welcomed them to the course as both speakers and listeners.

The announcement of the course is included below.

COURSE ANNOUNCEMENT: CS 412/EE 391

Purpose: To explore and exploit opportunities for automation of semiconductor manufacturing processes.

Topics: State of the art of semiconductor manufacturing automation in the lab and in industry. Increasing the effective use of computation in the design and control of manufacturing processes, through the combination of AI, computer graphics, and simulation. Designing the intelligent, interactive factory.

Format: An explicit goal of the course is to encourage cross-fertilization and collaboration between the fields of AI and semiconductor manufacturing. The course will include discussion of selected papers from both fields and participation in interdisciplinary team projects. Potential project topics include intelligent interactive interfaces for semiconductor processing equipment and the use of AI in design, simulation, monitoring, control, and diagnosis of semiconductor manufacturing processes.

2.7 New Electronic Discussion Group

To encourage and facilitate discussion of topics in computer-integrated manufacturing of integrated circuits, we have started an electronic discussion group called IC-CIM. An announcement of the mailing list is included below.

IC-CIM: A New Mailing List for Discussion of Computer-Integrated Manufacturing of Integrated Circuits

To join: Send your net address to IC-CIM-RequestSierra.Stanford.EDU.

IC-CIM is a new electronic mailing list for discussion of Computer-Integrated Manufacturing of Integrated Circuits. IC-CIM is maintained and moderated at Stanford University, within the Center for Integrated Systems.

The list of addresses for IC-CIM was initialized from lists supplied by Dave Hodges at Berkeley, Andrzej Strojwas at CMU, Paul Penfield at MIT, and Byron Davies at Stanford. A number of

industrial researchers have also been added to the list. New participants are welcomed from both university and industry.

IC-CIM is a moderated mailing list. When messages are sent to IC-CIM, they are gathered into digests of three or four messages and sent out this way to the mailing list. Readers will not be bothered by misdirected, inappropriate, or too frequent messages.

IC-CIM is open to discussion of any topic related to computer-integrated manufacturing for integrated circuits. Examples of IC-CIM topics are:

System architectures: computer hardware, software, networks AI and expert systems for semiconductor manufacturing

Processing equipment: capabilities, user interfaces fabrication simulation, scheduling and optimization, process and equipment modeling, process specification and design systems, manufacturing databases Data and knowledge representations for:

1. the semiconductor manufacturing line
2. fabrication equipment
3. wafers at each stage of fabrication
4. other fabrication materials (e.g., gases)

Training aids for operators, supervisors, engineers Integration of manufacturing, design and test systems

Chapter 3

Factory Modeling and Management

Michael Harrison, Avi Mandelbaum, Hau Lee, Avi Mandelbaum, Ann van Ackere, Lawrence Wein, Hong Chen, Anne Spence and Joe Calabrese.

Most of the research to be described in later sections of this proposal involves manufacturing productivity at the level of individual wafers. In this project, however, the focus is on factory-level productivity relationships, and more specifically on what might be called the throughput problem in semiconductor wafer fabrication. Manufacturing cycle time, also called throughput time, turnaround time, or the manufacturing interval, is typically five to ten times greater than total processing time in wafer fab operations, although the overall throughput rate (expressed in wafer starts per week, for example) may be substantially less than nominal system capacity. That is, wafers spend much more time queuing than being processed, even though critical equipment may be idle a significant fraction of the time. It is the desire to understand and ameliorate this phenomenon that motivates our research on factory modeling and management.

The subsections that follow describe three separate but related research tasks, each being undertaken with an industrial collaborator. Papers describing progress on the first two tasks have recently been completed, and these will be distributed to SRC member companies and submitted for publication shortly. Copies of the papers are included as attachments to this report, so the first two subsections below give only a brief overview of the corresponding attachment. The third task, which began this past summer, is still in progress, so the corresponding subsection gives a more lengthy description of work to date and plans for completion of the undertaking.

3.1 Predictive Performance Modeling for Wafer Fabrication

In this task we have been concerned with the use of queuing network models for analysis of wafer fabrication facilities. These rough-cut system models allow one to predict important performance characteristics of a wafer fab, such as the average throughput rate, average throughput time, or average work-in-process inventory, using only parameter values that are known at the design stage, or at least can be reasonably estimated.

In the attached paper, "Queuing Network Models of Semiconductor Wafer Fabrication", the congestion problems that plague wafer fabrication facilities are described in general terms, and several years' operating data from one particular facility are summarized. A simple queuing network

model of that facility is constructed, and the model is used to predict certain key system performance measures. The values predicted by the model are found to be within about 10% of those actually observed. Although a simple queuing model performs well for the one wafer fab we have studied in detail (at least for prediction of highly aggregated, system-level performance characteristics), we discuss refinements and extensions of the model that are likely to be important in other settings. This study, done in collaboration with the Hewlett-Packard Technology Research Center, involved one graduate student and two faculty members from the Graduate School of Business, and two graduate students from the School of Engineering.

The results reported in our paper suggest that queuing network models can provide useful quantitative guidance to designers of wafer fabrication facilities, allowing them to quickly and cheaply estimate the performance characteristics of various alternative configurations. Perhaps more important, however, is the simple qualitative point that congestion and delay in wafer fabrication are caused by variability in the operating environment. To significantly reduce manufacturing cycle times, one must reduce that variability.

3.2 Scheduling Wafer Fabrication

In this task we have been concerned with assessing the impact that scheduling can have on the performance of wafer fabrication facilities. The performance measure considered here is the mean throughput time for lots of wafers. In the attached paper, "Scheduling Semiconductor Wafer Fabrication", a variety of input control and sequencing rules are evaluated using a simulation model of a representative but fictitious fab. Certain of these scheduling rules are derived from a new body of theory developed by members of the project team, and their performance compares favorably with that of rules most frequently discussed in the scheduling literature. Three versions of the simulation model are used, which differ only by the number of servers present at particular stations, and which have one, two and four stations, respectively, that are heavily utilized (near 90% utilization).

The simulation results reported in our paper indicate that scheduling has a significant impact on average throughput time, with larger improvements coming from discretionary input control than from lot sequencing. The effects that specific sequencing rules have are highly dependent upon both the type of input control used and the number of bottleneck stations in the fab. This study, which also was undertaken with help from the Hewlett-Packard Technology Research Center, was done by Lawrence M. Wein, a graduate student in the Department of Operations Research (School of Engineering), and supervised by Professor J. Michael Harrison of the Graduate School of Business. Wein will continue his dissertation research on the theoretical aspects of scheduling, where a great deal remains to be done.

3.3 Lot Sizing and Rework Policy in Wafer Fabrication

In this task we consider the question of how lot sizing and rework policy affect the productivity of a wafer fab operation. The specific productivity measures considered are the throughput rate, manufacturing cycle time or turnaround time, and work-in-process inventory. Effort is initially being focused on modeling and analysis of the photo- lithography work cell, motivated by the following considerations:

1. Photo-lithography is generally considered to be the most complex and delicate operation in wafer fabrication.

2. Equipment involved in photo-lithography is usually very expensive. Hence, efficient utilization of the equipment by improved operating policies can be very beneficial.

3. The photo-lithography work cell is usually the bottleneck of the fabrication process.

4. Since wafers usually have to revisit the photo-lithography work cell several times for masking, wafers to be processed at the photo cell can be in very different stages of the fabrication process. This suggests that rework policies that are fabrication stage-dependent can be desirable. Moreover, the probability of a wafer lot having to be reworked is also a function of the lot size. Hence, the photo cell offers a lot of opportunities for productivity improvement by modifications in operating policies.

3.3.1 Specific Objectives

1. Lot-size control. Given an overall input rate and product mix, how would different lot sizes affect the rework rate and consequently the cycle time? What is the optimal lot size that minimizes average cycle time?

2. Rework policies. Given a rework policy, which may specify the rules for scrapping a wafer lot or reworking the defective ones, what would be the resulting cycle time, and what would be the necessary wafer start rate so as to achieve some desirable output rate under such a policy? What is the optimal rework policy that is desirable in terms of the average cycle time and wafer start rate or output rate?

3. Impact analysis of system improvement. What are the benefits resulting from improvements in set up times, yield variations, and rework rates?

3.3.2 Methodology

To investigate the above issues in the context of the semiconductor industry, it is essential to work with data from real wafer fabs and to interact closely with the management of those facilities. Motorola, Inc. has agreed to and begun collaboration with the Stanford research team on this project. Specifically, the fabrication operations at Mesa, Arizona, provide the setting for our study. This facility uses bipolar technology, and data for both a research lab and several production fabs are available. Essential elements of our research task are the following:

1. Understanding of system configuration, modular design, material flow, process and equipment characteristics, and operational control rules of the current facility. Data collection on output performance measures, process and equipment characteristics, demand and product mix requirements, and so forth. Identification of options for improvement, both in system design and operational control.

2. Development of a computer simulation of the photo-lithography work cell at the study site. The simulation model can serve several purposes: it can be used to provide quick answers to the first order effects of varying capacity and other resources at Motorola; it will be a

means for the project team to develop a simplified conceptual framework representing the flows of work and material at the work cell; it will be used to test the adequacy and accuracy of analytical models developed to address the managerial decisions described above; and it will also provide Motorola a vehicle to provide answers to other "what- if" questions for the evaluation of alternative design and operation policies.

3. Development of analytical models to study productivity relationships, and the determination of optimal batch size and rework policy. This task may involve the use of mathematical models such as queuing network systems and nonlinear optimization methodologies. The validity of the analytical model will be tested using the simulation model developed.

3.3.3 Task Personnel and Progress to Date

This task is a collaboration between Stanford University and Motorola, Inc., with Motorola personnel providing technical assistance with regard to process technology and the system environment, as well as feedback to the Stanford team. Mr. Doug Welter, manufacturing manager, is the contact person and project coordinator from the Motorola side. The Stanford project team is led by Hau L. Lee, Associate Professor of Industrial Engineering and Engineering Management, with Ann Spence, Ph.D. student in the Graduate School of Business, and Joel Calabrese, Ph.D. student in Industrial Engineering and Engineering Management. Professors J. Michael Harrison, Avi Mandelbaum, and Evan Porteus serve as consultants to the project.

Work on the task began with Ann Spence, a doctoral student on the project team, spending two months in July and August, 1986, as a summer intern at Motorola, under the direct supervision of Doug Welter. The purpose of Ann's mission is to work on research tasks (1) and (2) above. The on-site experience has enabled Ann to develop a good understanding of the complex system for wafer fab at the bipolar processing centers at Motorola. This understanding involves not only the photo-lithography work cell, but also other parts of the fab and their interrelationships with the photo cell. Initial data collection efforts resulted in the discovery of the non-existence or inaccuracy of certain key data. Data collection instruments have been devised and implemented, resulting in a preliminary set of data useful for the building of simulation for research task (2).

During the summer of 1986, a computer simulation model of the photo- lithography work cell has also been built in SIMAN. The details of this simulation model and the analyses performed using it are presented in a separate section. Results of the analyses have been presented to the management of Motorola, and initial reactions from Motorola have been very positive. It was felt that other production facilities at Motorola can benefit from similar analyses, and suggestions to repeat similar analyses at other sites were made. Ann Spence and Doug Welter are collaborating in a paper on the description of these analyses. The plan is to present the paper at the 1987 IEEE International Conference on Robotics and Automation, March 30 to April 3, Raleigh, North Carolina.

3.3.4 A Simulation Model

The simulation model is built to represent the photo-lithography work cell at the Bipolar Technology Center (BTC) of Motorola. Although mainly a research oriented facility, BTC also has a significant portion of efforts devoted to production needs. The simulation model, however, is expected to be representative of photo-lithography work cells at other production facilities. The simulation

language used was SIMAN, which was designed especially for use in simulating manufacturing systems. Output measures like the mean and standard deviation of cycle times, throughout rates, and other statistics on queue lengths and utilization of the resources, are recorded.

Different equipment for various processes like coating, baking, alignment, developing, and inspection are explicitly represented in the simulation model. Special attention is given to the distinction of fixed set up times, which are independent of the wafer lot size, and variable processing times, which are functions of the wafer lot size. This distinction enables an accurate evaluation of the impact of lot sizes on various productivity measures. At the inspection step, the simulation is currently built to reflect the actual operating procedures at Motorola. For a given lot, if every wafer in the lot passes inspection, the the lot exits the photo cell. If at least one of the wafer requires rework, the good wafers are then put aside until the reworks are completed, i.e., a lot always moves as a lot from cell to cell. Wafers requiring rework are to be resist cleaned in an adjacent cell, and then returned to the photo cell to be processed again.

A feature of the current simulation model is the explicit representation of operators as a resource at the photo cell. Operators are required to load and unload the lots for the vapor prime, and coat and bake track steps. They have also to be present for the alignment, developing and inspection steps. Hence, one can view the processing of wafers at the photo cell as services requiring both the equipment and operator resources. Queues will develop when either equipment or operators are unavailable. The system is also subject to random breakdowns of equipment, random repair times for breakdowns, and random breaks taken by operators. These events are also included in the simulation model.

Data on processing times, rework probabilities, test wafers, equipment breakdown frequencies, repair times, and operator breaktimes were procured from a variety of sources including specification manuals, data records and logbooks, personal observations and interviews with engineers and operators.

The current simulation also models work scheduling according to the existing operating procedures at BTC. Rework is given priority on the equipment, otherwise, the rule of first come first served is observed. In the case of operators, the priority rule of shortest remaining processing time is used to guide the operators on the next job they are supposed to perform on.

Experiments have been run on the simulation model, varying resource capacities, process capabilities, and operational rules. The performance measures include the throughput rates and cycle times. The results are summarized below.

1. As expected, adding more operators moves the tradeoff curve between throughput rate and cycle time rightward, although there are signs of decreasing returns to scale. Similar observation holds for adding more aligners. The simulation helps to assess the magnitudes of gains resulted from additions of such resources.

2. Hypothetical cases where the mean processing times and the probability of rework are reduced, are also tested using the simulation. The results show the magnitudes of gains in cycle time and throughput rate as a function of the degree of process improvements such as processing times and rework rates.

3. The standard lot size used at BTC is 20. Experiments with lot sizes of 10 and 40 are also run. The general result is that, for small lot size, the more frequent set ups required leads to longer cycle times for the same throughput rate. For very large lot size, however, the wafers

in a lot have to wait for all the wafers to complete processing before moving on to the next step. Such waiting is both a consequence of actual processing times on the wafers as well as a higher probability that rework will be needed for wafers in the lot. Hence, longer cycle time is resulted for the same throughput rate. The simulation results suggest that there exists an optimal lot size for the performance measures considered.

The analysis based on the simulation model described above is just a preliminary step in the overall research project. Nevertheless, it serves as an exercise for Ann Spence to familiarize herself with the fab conditions at Motorola, and a starting point for the collaborative efforts between Motorola and Stanford. Although still crude in nature, the results provide insights to some of the directions and magnitudes of the capacities and performance of the BTC facility at Motorola.

3.3.5 Future Plans

In the coming year, research will proceed along the following two fronts:

1. The simulation model will continue to be refined. Continual efforts for better data collection will enable the use of more appropriate assumptions on some of the distributions of variables used in the simulation. Our ultimate goal is to establish the simulation as a vehicle that is accurate enough to serve as a yardstick to evaluate analytical models as well as operational policies.

2. Work has just begun to formulate analytical models that represent the photo-lithography work cell. Our current thoughts are that these could take the form of queuing network models with multiple classes of customers. The classes of customers correspond to the stage of the wafer lot in terms of the fabrication process (e.g., the number of masks that have been completed), and the rework size. Based on such models, the impact of varying lot sizes and rework policies can be explored.

Chapter 4

Simulation of Processes, Devices, and Circuits

4.1 High-Level Process Simulation for VLSIC Fabrication

Steven D. Leeke and John Shott

4.1.1 Introduction

The purpose of this research is to simulate a VLSIC process using physical process and *equipment* models in such a way as to facilitate statistically significant simulation results. This requires high-volume process simulation, or what can be called **manufacturing-level process simulation**. The simulations are performed with a new process simulation system, **SHIPS** (Stanford High-Level Incremental Process Simulation) that improves on past work by utilizing incremental simulation, a compact simulation language that is structured according to a set of abstractions that are familiar to processing personnel, and a menu-driven, form-based user interface that provides an intelligent framework for creating new processes and simulations. Furthermore, the system is designed to perform the simulations in a simulation *engine* independent manner. For the work presented here the advanced process simulation capabilities of the **SUPREM** family of process simulators is used as the simulation *engine*.

4.1.2 Previous Efforts

To date there have been few efforts at manufacturing-level process simulation. Some of the better known and successful efforts have come from CMU and Hitachi, with the **Fabrics** family of programs and the **CASTAM** program, respectively[4.1.1,4.1.2,4.1.3]. The former is an evolving set of tools for process information capture and simulation, device simulation, and parameter extraction for circuit simulation. The **Fabrics** process simulation is done using only analytic models. This has the advantage of speed, but the disadvantage of large errors in both process and device simulation results for state-of-the-art technologies.

The **CASTAM** work is a more thorough attempt at statistically significant full-process simulation. One of its fundamental assumptions, however, is that whatever input parameter distribution, via Monte Carlo simulation, produces the correct output distribution with respect to the measured

values, correctly represents the actual input parameter variations. Their results were good, but they provided little, if any, physical insight into the causes of the variations in the input parameters.

4.1.3 Proposed Improvements

The work being described here will improve on previous work by:

- incorporating physical equipment models, as they evolve, for producing meaningful input parameter distributions.
- using the advanced physical process models of the **SUPREM** family of process simulation programs.[4.1.4,4.1.5,4.1.6,4.1.7]
- utilizing an incremental simulation algorithm to reduce the overall CPU time required for a simulation.
- utilizing a high-level process simulation language, **SHIPS**, that provides a very compact way to represent complex process simulations and interface to device simulation programs.
- including a three level abstraction of VLSI processing to reduce the complexity at any one level. The three levels are, in descending order, *process*, *module*, and *step*.
- providing a structured user interface that is menu-driven, uses form-based entry, and includes extensive on-line help. This will reduce the level of expertise required to profitably utilize the system.
- using the concept of a *process module library* to augment the processing abstraction. This provides a structured way to support many different technologies simultaneously.

While the issue of integration of process, device, and circuit simulation tools has been addressed in recent publications,[4.1.8,4.1.9] and will only become more important, those issues are beyond the scope of this work.

4.1.4 The SHIPS System

The **SHIPS** system consists of an integrated set of components. These components are:

- A menu-driven and form-based user interface for creating processes, modules, and steps.
- A language, **SHIPS**, for describing multi-parameter variations in a process in a structured and compact form.
- A compiler for efficiently generating and supervising the process simulations described with the **SHIPS** language.

These components are centered around the **SHIPS** compiler. This compiler uses the high-level process simulation description language **SHIPS**. The compiler and the language depend on the abstractions used in the methodology of a *process*, *module*, and *step*. To increase the user's efficiency in creating processes, modules, and steps a form-based and hierarchical menu-driven user interface is used.

4.1.5 The Architecture of the SHIPS System

The SHIPS software is being written in C++, an object-oriented version of the C programming language by AT&T. The first two prototypes of the software were written in C, but the decision was made to move to C++ for the final version. The fundamental reason for this change is extensibility.

Using C the program's architecture was based on control-flow programming, that is, the control constructs of the program determined how the data was manipulated. In the C++ version, however, the object paradigm supports logic-flow programming, or the data itself determines how it is manipulated according to interactive events. This difference can be summarized by noting that in control-flow programming the paths through the program can be determined *a priori*, while in logic-flow programming the paths through the program are determined by events.

This logic-flow approach is crucial to the extensibility of the SHIPS system. This approach allows a specification to be developed for adding new processing steps and models and equipment functions to the program in a simple way. In addition the modularity of the program is increased dramatically as all operations on an object are defined by that object.

4.1.6 The SHIPS High-Level Simulation Language

The SHIPS language is a structured language designed to provide a compact way to represent large-volume process simulations in a simulation *engine* independent manner. The SHIPS compiler assumes the following about the simulator being used: that it can read and write intermediate files and simulate a complete VLSI process.

The language has constructs for defining the *process* to use, the *module library*, and the *views* of the process to be simulated. A *view* is, in the simplest sense, a certain process sequence as determined by the lithography of the process. For example, the channel view and the source-drain view are different. For a one-dimensional simulator such as SUPREM III the natural view is a profile, while for a two-dimensional simulator it becomes a cross-section.

The language incorporates numeric expressions, functions, and variables; string variables; a block structure based on the modules of a process; for-to and set assignment statements for multi-value assignments; simple statistical and factorial analysis of the simulation results; and integration to other simulation programs for post-processing of the process simulation results via output statements and *yield parameters.*

An Example of the SHIPS Language

The following is an example of a simulation program written in the SHIPS language. The keywords of the language are in boldface type. An explanation of the effects of the program are included as comments in the example.

```
// The SIMULATION statements is merely a way to start the program and
// any text that appears after the keyword and before the semicolon is
// included in the output as documentation. Carriage returns are preserved
// in this text.

SIMULATION Advanced Example;
PROCESS su2uCMOS(pch,nch,dnch,nsd,psd,field);
```

```
// su2uCMOS is the name of the process to be used. The parameters
// are the names of the predefined views of the structure to be
// simulated.

{
    This is a comment. It must be surrounded by brackets.
    It can contain any amount of text, including carriage returns
    and appear anywhere appropriate in the text.
}

// This is also a comment. Anything that appears after // until the end of
// the line is considered a comment.

// Below are all of the variables and their initial values. String and
// numeric variables are both supported. The initial appearance of
// a variable must be outsize of a WITH block.

base_time = 100;
step_value = 10;
base_energy = 110;
base_dose = 1e10;
process1 = FALSE; // FALSE is a predefined value.

// Conditional statements using C syntax are supported.

IF ((base_time < step_value) || (step_value == 10)) THEN

    // The WITH statement is used to modify the parameters
    // in the module that it specifies. WITH statements may
    // not be nested.

    // NOTE: WITH statements may appear in any order in the
    // program. The do not have to appear in the order they
    // occur in the process flow.

    WITH nit1 DO // nit1 is a process module
        IF (base_dose == 1e12) THEN
            thickness = 0.1000; // The thickness parameter is
        ELSE // changed depending on the
            thickness = 0.25; // base_dose variable value
    ENDELSE;
    END;
ELSE
    IF [illegible] alue == SQRT(base_time)) THEN
```

```
        WITH nitride_mod DO

            // A set statement is used to assign multiple
            // values to a process module parameter.
            // A simulation will be run for each of these values.

            SET thickness = [120,125,150];
        END;
    ENDIF;
ENDELSE;

IF process1 THEN
    WITH kox DO

        // FOR statements are used to automate multiple
        // parameter value assignments. Increments in a FOR
        // statement can be additive or multiplicative.

        FOR time = 100 TO 1000 STEP MULTIPLY 10;
    END;
ENDIF;

WITH psdi DO

    // When the STEP keyword and modifier are not supplied
    // the an increment of one is used.

    FOR energy = base_energy to base_energy + step_value;

    // A full range of functions are supported, although user defined
    // functions are not. These functions may appear anywhere a numeric
    // value is needed.

    SET dose = [EXP(14),1e10,1e11,1e12,5*10^LOG(1e12),10^ln(1e14),5e14];
END;

WITH nwi DO
    // Statements are delimited by semicolons, not lines.
    energy = 120; dose = 1e15/SQRT(base_dose * base_time);
END;

// The OUTPUT statement indicates what information should be extracted
// from the fundamental simulator that is being used, and how it should
// be tabulated.
```

OUTPUT jctn_depth mean variance std_deviation;

// The **SIMULATE** statement indicates what filename to use for the output,
// adv_example in this case, and what module to start with (init) and what
// module to end with (psg). Additional options include the ability to
// specify an intermediate simulation file to use to start the simulation.

SIMULATE adv_example **FROM** init **TO** psg;

4.1.7 Incremental Simulation

Simulating the effect of the variations in certain process parameters can become an imposing task when either the complexity of the process is high, the number of parameters being varied is large, or the number of values assigned to any one parameter is large. At present, most simulations are done in an *a priori* fashion in that the *entire* simulation is repeated each time a parameter is changed. Incremental simulation utilizes the modularity of the process description to reduce the complete simulation to the minimum number of modules possible. The simulations are generated such that *only* those modules which will be impacted by a particular parameter change will require re-simulation (with intermediate storage of the results of each incremental simulation). Keeping track of these intermediate simulations can clearly become a sizable bookkeeping task. This requires a simulation system that can control the incremental simulation process.

In summary, incremental simulation allows the total simulation process to be reduced to the minimum number of module simulations. This shows a dramatic improvement over *a priori* simulation generation. The following definitions will be used in the expressions for the number of modules simulated during normal *a priori* and incremental simulation.

- $M \equiv$ The length of the process in modules.
- $I \equiv$ The number of modules that were changed in the **SHIPS** program such that the parameters of that module have a product of changes that is greater than one. This can be expressed as the requirement that $\prod_{p=1,P_k} V_{pk} > 1$ for the k^{th} module in the process.
- $P_i \equiv$ The number of parameters assigned two or more values in the ith module that was changed.
- $V_{pi} \equiv$ The number of variations of the pth parameter in the ith module that was changed.
- $\delta m_i \equiv$ The number of modules simulated in the ith level of the incremental simulation. In the case of a module containing only one step, $\delta m_i = 1$, and the equations that follow simplify.
- $S_{old} \equiv$ The number of modules simulated by creating all of the simulation input programs *a priori*.
- $S_{new} \equiv$ The number of modules simulated by doing the simulations incrementally, i.e. simulating one module change at a time.

Module Number	Parameters Changed	Number of Changes for Each Parameter
1	1	3
10	1	4
20	2	5,4
30	3	5,2,3
40	2	3,3
50	4	2,2,3,4

Table 4.1.1: Incremental simulation algorithm example for M = 60.

Given these definitions the following expressions can be written for the total number of modules simulated with both the old approach and the incremental simulation approach.

$$\begin{aligned} S_{old} &= M\left[\prod_{i=1,I}\left(\prod_{p=1,P_i} V_{pi}\right)\right] \\ &= \sum_{i=1,I}\left[\left(\prod_{j=1,I}\left(\prod_{p=1,P_j} V_{pj}\right)\right)\delta m_i\right] \\ S_{new} &= \sum_{i=1,I}\left[\left(\prod_{j=1,i}\left(\prod_{p=1,P_j} V_{pj}\right)\right)\delta m_i\right] \end{aligned}$$

Making the following definitions:

$$\begin{aligned} A_i &= \prod_{j=1,i}\left(\prod_{p=1,P_j} V_{pj}\right) \\ B_i &= \prod_{j=i+1,I}\left(\prod_{p=1,P_j} V_{pj}\right) \\ S_{old}/S_{new} &= \frac{\sum_{i=1,I} A_i B_i \delta m_i}{\sum_{i=1,I} A_i \delta m_i} \end{aligned}$$

This implies that $\max(B_i) \geq S_{old}/S_{new} \geq \min(B_i)$ and in general $S_{new} < S_{old}$ for $I > 1$ and $S_{old} = S_{new}$ for $I = 1$.

Incremental Simulation Example

For a process with 60 modules, Table 1 shows the modules that have changed parameters and the number of changes. Only those modules that have had parameter changes are shown. The simulation results come from a program that simulates the algorithm.

Simulated results for the simulations are as follows:

$$
\begin{aligned}
S_{old} &= 186,624,000 \text{ modules simulated} \\
S_{new} &= 34,936,947 \text{ modules simulated} \\
\frac{S_{old}}{S_{new}} &= 5.34 \text{ and} \\
\frac{S_{new}}{S_{old}} &= 0.18
\end{aligned}
$$

This shows that new incremental simulation algorithm reduces the number of modules to be simulated by over 400%. While the number of modules is still large, the reduction is significant. In addition, this example points out how expensive it is run this kind of simulation. The solution to this problem is the application of massively parallel computers to the problem.

The incremental simulation algorithm lends itself well to parallel machines as it becomes increasingly a parallel problem for cases such as the example above. For those problems, the figure of interest is not the number of modules to be simulated but the height of the tree in modules, M, and the total number of steps to be simulated in each of the modules. From these numbers and the parameter variations we can compare the effectiveness of the two approaches.

The number of processors needed in order for the *a priori* method to compete with the incremental simulation algorithm is the full width of the tree, S_{old}/M, since each variations requires a full process simulation. With the incremental algorithm the advantages in parallel operation are that the complete problem can be started with very few process and given a maximum number of processors, P, the complete problem may be worked on in parallel up to a point where the number of different simulations in progress is $S_{old}/(MP)$. When that point is reached the different simulations will require sequencing as the the processors become available. Hence, portions of the tree can be worked on and the gains over the *a priori* approach in serial processing again become apparent.

4.1.8 Equipment Functions

The problem of statistical process characterization has been defined by Spanos and Director[4.1.10] as determining the statistical moments that define the distributions of the process fluctuations, once the moments of the measured process outputs are known. The problem was formulated in their paper in terms of determining the nominal and random components of the process and device parameters. The nominal parameters of a process, however, can be determined specifically based on equipment models, rather than assuming that they are determined from measurements only.

The process environment that any wafer sees is determined by the processing equipment. The equipment establishes the environment based on several factors. These factors include both specific and random variations. Examples of specific variations are those introduced either by specific control or by specific design of the equipment. Random variations are those that can not be attributed to a specific cause. Hence, specific variations in the process environment can be determined by modeling the equipment.

There are two fundamental aspects to an equipment model. First, determining the input parameters a particular equipment function[1] requires. This amounts to the task of defining the fundamental control parameters of the equipment. Second, determining the output values of a particular

[1] An equipment function is the implementation of an equipment model.

equipment function. This amounts to the task of defining what parameters of the equipment's environment should be extracted.

The purpose of the equipment function is not to do any process simulation, but rather to simply return a set of parameters that describe the process environment. Preliminary versions can simply be distribution functions relating an input configuration to an environmental parameter distribution.

4.1.9 Statistics

The SHIPS system is designed to analyze the results of a simulation based on simple statistics. These include the use of:

- mean
- standard deviation
- variance
- median
- maximum
- minimum
- range
- skewness
- kurtosis
- factorial design analysis
- n-way ANOVA
- multiple regression

While most of the above is self-evident in its applications, two merit further discussion, n-way ANOVA analysis and multiple regression.

The n-way ANOVA analysis allows the most significant input parameters to the process to be identified for *each* output parameters (junction depth, threshold, sheet rho, etc.). The multiple regressions allows the most significant parameters, as identified by the n-way ANOVA analysis, to be fit to the data such that analytical expressions can be derived that describe a process accurately. These provide fast analytic feedback on how the process is performing. In addition, with further analysis, as will be explained in the next section, the process sensitivites can also be studied in analytic form.

4.1.10 Applications of Artificial Intelligence

Artificial intelligence workstations, such as the the Texas Instruments Explorer[4.1.11], are now available on campus and they present unparalleled opportunities for enhancing high-level process simulation in a number of ways. We will elaborate on only one at this time.

The analytical analysis described above is valuable - but only to a limited extent. To be able to describe a process is a closed-form analytic manner gives process development and manufacturing engineers a power set of tools to use. In addition, however, the ultimate goal, particularly for a manufacturing engineer is to develop a *manufacturable* process. This requires that robustness be built into the process. To accomplish this the sensitivities of the process must be explored and ways to improve the sensitivity performance of a process developed.

The process sensitivities can be studied in analytic form using the the following approach. Given the analytic expressions for a process' structure or electrical performance in terms of process inputs (as a result of the n-way ANOVA analysis) an AI tool called MACSYMA[4.1.12] is capable of producing the derivatives of those expressions - in *symbolic* form. This means that the first and second derivatives of a process description can be used to accurately understand the sensitivity performance of a process.

4.1.11 Activites During the Past Year

During the past year the following was accomplished

- The incremental simulation algorithm was developed to reduce simulation costs.
- Three version of the language were written - each time as a direct result of input from computer science researchers and processing needs.
- A prototype menu-driven user interface was developed using the Unix C shell. A second was written in C. The user-interface was not completed and released as future needs pointed to a need for more extensibility than could be achieved with C.
- C++, an object-oriented version of C, was evaluated for the release implementation.
- Two prototype compilers were written in C. Due to the changes in the user-interface the compiler work was halted until the implementation language and environment are selected.

4.1.12 Future Plans

The near-term goals (6 – 18 months) for the SHIPS system are:

1. Evaluate artificial intelligence tools for their application to IILPS.
2. Evaluate the Lisp machine environment for the release implementation of the SHIPS system
3. Implement the complete SHIPS system in the best environment/language possible
4. Complete a library of process modules for the Stanford 2 μm CMOS process
5. Perform statistical simulations of the Stanford 2 μm CMOS process and compare the results to measured data
6. Evaluate the possible addition of PREDICT from MCNC as a fundamental process simulator

4.1.13 Conclusion

The **SHIPS** System for high-level process simulation provides a high-level process description language, a structured user-interface, and a new simulation algorithm for reducing the number of modules that must be simulated in large volume simulations. The proper implementation vehicle is being determined and following that the compiler and user interface will both be implemented.

Our implementation incorporates the advanced physical process models of the **SUPREM** family of process simulation programs in conjunction with physical equipment models to provide physically meaningful input and output process parameter distributions. The **SHIPS** system, however, is designed to be simulation engine independent.

4.1.14 References

[4.1.1] Shojiro Asai Yukio Aoki, Toru Toyabe and Takaaki Hagiwara. Castam: a process variation analysis simulator for mos lsi's. *IEEE Transactions on Electron Devices*, ED-31(10):1462–1467, October 1984.

[4.1.2] Andrzej J. Strojwas Sani R. Nassif and Stephen W. Director. Fabrics ii: a statistically based ic fabrication process simulator. *IEEE Transactions on Computer-Aided Design of Integrated Circuits*, CAD-3(1):40–46, January 1984.

[4.1.3] Wojciech Maly and Andrzej J. Strojwas. Statistical simulation of the ic manufacturing process. *IEEE Transactions on Computer-Aided Design of Integrated Circuits*, CAD-1(3):120–131, July 1982.

[4.1.4] S.E. Hansen C.P. Ho, J.D. Plummer and R.W. Dutton. Vlsi process modeling - suprem iii. *IEEE Transactions on Electron Devices*, ED-30(11):1438–1453, November 1983.

[4.1.5] K. Doganis L. Mei R.W. Dutton, P. Fahey and H.G. Lee. *Computer-Aided Process Modeling for Design and Process Control.* Special Technical Publication 804, American Society for Testing and Materials, 1984.

[4.1.6] R.W. Dutton. Modeling of the silicon integrated-circuit design and manufacturing process. *IEEE Transactions on Electron Devices*, ED-30(9):968–986, September 1983.

[4.1.7] Robert W. Dutton and Stephen E. Hansen. Process modeling of integrated circuit device technology. *Proceedings of the IEEE*, 69(10):1305–1320, October 1981.

[4.1.8] Duane S. Boning and Dimitri A. Antoniadis. Mastif - a workstation approach to fabrication process design. In *Proceedings of the ICCAD*, pages 280–282, IEEE, Santa Clara, CA, November 1985.

[4.1.9] Ralph J. Sokel and Donald B. Macmillen. Practical integration of process, device, and circuit simulation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits*, CAD-4(4):554–560, October 1985.

[4.1.10] Costas J. B. Spanos and Stephen W. Director. Parameter extraction for statistical ic process characterization. *IEEE Transactions on Computer-Aided Design of Integrated Circuits*, CAD-5(1):66–78, January 1986.

[4.1.11] *Explorer Technical Summary.* Texas Instruments Incorporated, Data Systems Group, P.O. Box 2909 M/S 2151, Austin TX, 78769, 1985.

[4.1.12] *Macsyma Reference Manual.* The Mathlab Group, Laboratory for Computer Science, MIT, Cambridge, MA, second printing edition, December 1978.

4.2 Compilers for Improved Device/Circuit Simulation

Faheem Akram and John Shott

During this period a major portion of the compiler for the above mentioned language was completed. This included the lexical analysis phase, the parsing phase, the symbol table generation, and the implementation of the error recovery mechanism. What was missing at this stage to make it a complete compiler was basically the semantic actions that are to be carried out once a correct sentence of the language has been recognized. These had been delayed pending a resolution of certain technical questions pertaining to not only the use of the language but also to whether the compiler should output a program in an intermediate language that may be interpreted or whether the compiler should output a program that is acceptable to the C compiler. The most efficient approach would be to generate the machine code directly. However the approach of generating C code has many advantages in the context of our research effort..... it allows us to complete the compiler in less time. As a result we can concentrate on the language and the features it supports and strive to make it more useful in a manufacturing environment. It also makes it easy to link in specialized C language programs with the code generated by our compiler. Thus while our language is much simpler than "C" it can potentially have the full power of "C" available to the more sophisticated user. From a manufacturing point of view this is very desirable, in that it allows the engineers on the manufacturing floor to learn the language quickly for their day to day use, while still providing for sufficient power that can be of use in specialized applications in the research lab associated with the manufacturing floor.

The language and the associated compiler were completed around the beginning of July. A program was written in our language for PISCES and it ran successfully. The experiment was to look at the effect of varying the substrate bias on the threshold voltage of a MOSFET transistor. It was found that the ability to couple in C language programs was a very powerful aid in writing programs that could also interface with the unix system. In particular we were able to benefit from the Makefile mechanism of unix in creating a program that only did new grid generation only when the source or drain doping profiles had changed. Thus unnecessary runs of a particular piece of code could be avoided. The experience was wholly satisfactory from the point of view of initial goals. It was then decided to apply our language to the automation of other simulators such as SPICE. Initially it was anticipated that changing the table of keywords that the compiler uses would be all that would be required. The keyword table not only contains the keywords of our language but also the keywords of the simulator to be incorporated. Thus by changing the keywords that belonged to PISCES to those that belong to SPICE we expected to be able to automate SPICE and for that matter any other simulator. However this did not turn out to be that easy. Our compiler had to be able to distinguish the text for a simulator (such as PISCES) from the constructs of our language. This was accomplished by looking, noting that in PISCES all its statements are begun by a keyword. Thus whenever our compiler came across a keyword of PISCES it would assume that from that word on to the next semicolon it was looking at PISCES text. With this approach we did not have to build a full-fledged parser for the language of the PISCES program. Also it was believed that most other simulators had their languages defined in such a way that each line of their text would begin with a keyword and hence to change our compiler to include a new simulator would involve only updating the keyword table of the simulator. This however did not turn out to

be the case for SPICE. In SPICE a statement defining the location of a resistor and its value is

Ra 10 20 1000

Now as part of our language one can have a statement of the form

Ra = Ra + 100;

Thus it becomes difficult to determine whether a particular line beginning with *R* as a keyword belongs to our language or to SPICE. This could be very easily handled by extending the grammar of our language a little bit. However this was deemed rather undesirable for our initial goal of extending our language to other simulators by merely updating the keyword table would not be met. Instead what would be required is a modification of the grammar for each new simulator. Thus a new approach was taken. The basic idea is to mark the beginning and end of a piece of simulator text by keywords @begin(simulator_name) and @end(simulator_name). This way whatever comes in between these to keywords is known to belong to the simulator whose name is simulator_name, and hence the compiler can distinguish program text from simulator text. Further our experience with the features of C language that were not included in our language but were available to us through the use of external C functions that could be used in programs written in our language convinced us that it would be better to extend the C language so that the ability to introduce variables in a simulator program is included in it. At this juncture I became aware of a new language called C++. It is an extension of the basic C language so as to allow object oriented programming. It was realized that such a language could reduce the job of writing the new compiler very drastically. Hence I have recently learnt C++ and now am working towards the goal of extending C++ so it can help incorporate variables in a given simulator. In addition we hope to be able to use different simulators in the same file. Thus one could write a PISCES program for MOSFET subthreshold characteristics in which one looked at the variation of Id versus Vgs at a small Vdd while varying the threshold adjust channel implant. Then one could call an RS1 routine from within the same program to extract the statistical information such as the mean and standard deviation of the threshold voltage. It still however will remain true that one will only have to know very little C++ to be able o effectively use our tool. However the more one knew about C++ the more one will be able to accomplish.

4.3 Statistical Process/Device Simulation

Shahin Sharifzadeh and John Shott

4.3.1 Introduction

With the progress made in the fabrication of semiconductor devices, VLSI circuits with near-micron channel lengths and high densities are being manufactured in large numbers. The goal of the VLSI process engineer is to provide the circuit designer with high quality transistors which show minimal variation of device characteristics. Computer simulations of the process and device are frequently used to calculate values for film thicknesses, doping profiles, and device characteristics. To estimate the sensitivity of the process or device to a particular parameter, the means and standard deviations of device characteristics can be predicted by appropriately varying process parameters in the simulation program. The process can then be optimize to obtain the highest achievable yield and the lowest device variations.

4.3.2 Simulation Using Make Program

SUPREM, a process modeling program developed at Stanford University was used to perform the process simulations. The entire 2 μ-CMOS process was divided into small modules. Each module represents a process step, i.e., Gate Oxidation, N-well Implant, etc. By breaking the process into small modules, a large amount of cpu simulation time can be saved. This is extremely important in a manufacturing environment where a large number of simulations are needed to obtain accurate statistical distributions for device characteristics. For example, consider the effect of temperature variations in the Gate Oxidation step on the resulting device characteristics. Since the process steps occurring prior to the Gate Oxidation will not observe such variations, the profiles and film thicknesses will remain unchanged. Obviously, the simulation can be started from the point of initial parameter variations and the total cpu time can be reduced drastically by performing incremental simulation of the process. In order for this technique to be applicable to a complex simulation, however, we must have means of automatically determining *precisely* which modules require resimulation to accommodate a particular process change

The number of modules increases as the process gets more complicated. The task of organizing, simulating, and saving the intermediate results will be enormous for any VLSI processes. To make the modular technique acceptable in a manufacturing environment, all steps related to simulation, saving and management of the modules should be transparent to the user. One of the features of the Unix Operating system which matches this purpose is the *make* program. *Make* is designed to control the compilation, assembly, and loading of complex assemblies of software. In particular, it recompiles *all of the files*, but *only those files*, which will be dependent on a change in a particular software module. *Makefiles* provide the means for describing the nature of the dependencies in and assembly of software routines (or, in this case, simulation modules) and the means for specifying the actions which are to be taken in the event that a module will be affected by a particular change. With the makefile structure available, modular simulation of a complex 2 μ-CMOS process can be performed on channel, source/drain, and field regions of both p- and n-channel transistors.

Once the makefile structure is in place the intermediate steps involved in the complete simulation will be transparent to the user. Information such as threshold voltage, sheet resistance, and I-V characteristics can be extracted automatically by the appropriate specification in the *makefile*. Since all the intermediate results are saved, device characteristics and various profiles are available after simulation of each processing step.

4.3.3 Factorial Design

Factorial design experiments can be used to extract information about effects of process inputs on device or circuit parameters. In general multi-factor experiments can be preformed to look at the changes of system output due to variations in the inputs. Analyzing the data of a factorial experiment yields main effects and interaction terms of the inputs. Inputs with large main effects have greater impact on system's output variation.

Factorial experiment is a powerful tool in understanding processing variations. The large number of processing steps in any VLSI circuit makes it difficult to predict effects of process variations on circuit performance. By organizing an appropriate factorial experiment, the role of process inputs on device and circuit parameters will be achieved quickly. Multi-factor experimental design requires a large number of experiments (depending on number of inputs and the level of the factorial design). Therefore simulating as opposed to measuring process and device characteristics would be cheaper and faster.

The *make* structure developed to simulate the 2μ-CMOS process which uses SUPREM and PISCES as simulators is used to preform the necessary simulation for the factorial experiment. The incremental simulation of the Makefile will reduce the cpu time to a minimum.

A simple 2 level factorial experiment with three variables was designed to look at the effects of N-Well implant, Boron Adjust implant and Channel Length on threshold voltage (table 1). In table 2, the main effects and interaction terms which were calculated using Yates algorithm can be observed.

4.3.4 Polynomial Regression

Statistical analysis of any parameters requires a large number of samp'es. Process and device simulations are very CPU intensive, therefore it is not prudent to use simulators to generate the necessary points. The solution proposed is to find a polynomial expression which relates the desired output to values of process parameters.

Finding an accurate polynomial fit is difficult because we are dealing with a large number of inputs. By elimination of less important parameters, the polynomial expression can be simplified. Exclusion of process parameters can be achieved by preforming the factorial experiment outlined in previous section. The main effect of each input can be used as a selection parameter.

The polynomial is used to predict the value of the desired output. Since we only have to compute the value of the polynomial for a given set of inputs instead of preforming full process and device simulation, the results can be calculated very rapidly. It must be noted that the results from polynomial expression is only valid in a certain region of inputs. Process parameters outside this range will result in invalid values. As it was mentioned, computation of polynomials are extremely fast; therefore analysis which otherwise would have been impossible (CPU consideration of process/device simulators) can be preformed. Distribution of output parameters can be obtained using Monte Carlo simulation.

4.3.5 Future Activities

The main objective of any process engineer will unquestionably be to design a *process for manufacturability*. In this activity we are exploring techniques which can be used to aid process engineers address and understand the problem of *process for manufacturability* effectively.

Efforts will be directed into finding the optimum scheme for preforming factorial experiments, namely number of levels and variables. Monte Carlo simulation is a very powerful technique for obtaining distribution for parameters. Since this method requires a large number of simulations, it is crucial to find the appropriate polynomial expression. In addition to Monte Carlo simulation, the polynomial expression can be used if analyzed issues on process control. Process control criterion will be examined and device/circuit parameter distribution would be obtained and compared to actual devices fabricated in the lab.

Table 4.3.1: 2^3 factorial design.

Experiment No.	NWI	BAI	ChL	V_t
1	-	-	-	-0.9010
2	-	-	+	-0.9492
3	-	+	-	-0.8687
4	-	+	+	-0.9080
5	+	-	-	-0.9211
6	+	-	+	-0.9574
7	+	+	-	-0.8977
8	+	+	+	-0.9459

	+	—
NWI	2.563E12	2.438E12
BAI	4.12E11	3.88E11
ChL	1.69	1.47

Table 4.3.2: Interaction terms for 2^3 factorial design.

Interaction	Value (Volts)
Avg.	-0.9186
1	-0.0430
2	0.0271
3	-0.0024
1-2	-0.0008
1-3	-0.0008
2-3	-0.0097
1-2-3	-0.0052

1 ..... ChL
2 ..... BAI
3 ..... NWI

Chapter 5

Technology - Equipment Modeling

In the Technology project we are developing a new modeling discipline-semiconductor manufacturing equipment modeling. Perhaps the greatest obstacle to simulating a semiconductor manufacturing line is an almost total lack of physical models to describe process-parameter variations resulting from characteristics (e.g. the geometry) of particular machine designs. Existing models used in process simulators such as SUPREM are highly generic and assume that a specific local environment is replicated across an entire wafer, from wafer-to-wafer in a batch and from batch-to-batch. This assumption is an invaluable simplification in developing process models per se but neglects *the* fundamental manufacturing problem of parameter distributions and their causes. Equipment modeling will address this problem. The Patterning project is concerned with modeling optical, electron beam and other lithography equipment. The Etching project is focussing on modeling plasma, reactive ion and other etching equipment. The Deposition and Redistribution project includes modeling the equipment for physical and chemical vapor deposition, oxidation and diffusion, rapid thermal processing, and ion implantation. The general approach to each project begins with the development of new test structures and measurement tools to improve understanding of manufacturing parameter distributions. This experimental data is then be used as a guide to the formulation of predictive physical models for various machines. Agreement between measured and theoretical distributions will verify the models and serve as a basis for defining in concert with manufacturers new machine concepts for future generations of equipment.

5.1 Kinetic Modeling of Active Species of Dry Etching

K. S. Uhm, M. R. Kump, J. P. McVittie, and R. W. Dutton

5.1.1 Introduction

Dry etching now plays a dominant role in the evolution of IC devices. As the density of devices increases beyond the current VLSI level, the control and understanding of dry etching techniques is critical to optimization of the fabrication processes. This requires, in addition to extensive empirical investigation, a dry etch model accounting for the fundamental physics and chemistry of RF glow discharges. However, due to the complexity of the reactions involved in dry etching processes, much

of the process development in this area has been based on empirical results with little support in the way of an engineering model to guide in the development of processes and equipment.

This chapter reports a simple but useful engineering model of dry etching which enables us to develop computer simulation to optimize the design of dry process technology. Dry etch processes depend on the complex interaction of ion bombardment[5.1.1] and the chemical reactions[5.1.2] of species. In this study the focus is on the modeling of the spatial distribution of important reactive species and volatile by-products based on simplified kinetic reactions and several basic assumptions. The vehicle for this study is Si etching using SF_6 + O_2 + Ar plasma in a parallel plate type etcher. By comparing the computer simulation results with the test data, the effects of basic plasma parameters, such as diffusion coefficients, lifetimes, and surface reaction coefficients, on dry etch processes are evaluated. This modeling of the active species, when combined with models for the ion bombardment and surface reactions, should yield a powerful tool for modeling dry etching processes.

5.1.2 Conceptual Model

In dry etching processes, there are many different kind of important plasma reactions, which include generation, recombination, and surface reactions of critical species. As mentioned already, the SF_6 + O_2 + Ar system was chosen to study these reactions, because, first, SF_6 is a good fluorine source without supplying any fluorocarbon chemicals, second, O_2 enhances the fluorine generation so that the system shows both high F concentration and strong F^* emission intensity[5.1.3], third, Ar is used for the "actinometer" method[5.1.4]. As a result, this system enables us to study the effects of photoresist presence to the silicon etching processes, because fluorocarbon chemicals are mainly produced as a by-products of the reactions between photoresist and fluorine atoms.

Out of many complicated interactions of this system, our studies are focused only on the generation of fluorine atoms and their reactions with the silicon substrate and photoresist, since fluorine atoms play the dominant role in the etching processes. For the further simplification, among many species which are related to fluorine atoms in this system,

four species (Ar, F, SiF_4, and CF_2) are chosen to be measured and to be included in the dry etching model.

A conceptual model which summarizes these plasma kinetics is shown in Figure 5.1.1. From this concepture model, it is clear that fluorine atoms react with the silicon substrate and photoresist layer to form SiF_4 and CF_2, respectively, along with many other by-products such as CO_2 and etc.. These by-product species will diffuse away from the reaction surfaces and will, eventually, fall onto the substrate or be involved in loss mechanisms due to the recombination reactions and the effect of vacuum pumping.

5.1.3 Dry Etch Model

Based on the experimental results and the fluorine generation model of this study as reported in [5.1.5], a simple quantitative kinetic model of the SF_6 + O_2 + Ar system can be constructed as listed in the following relations by selecting only the dominant chemical and physical reactions related to fluorine species from the many possible reactions[5.1.3] occurring within the plasma system.

Bulk Generation Reaction Equations :

$$SF_x + e^- \xrightarrow{k_{Fx}} SF_{x-1}^- + F \quad (x = 4,5,6) \tag{5.1.1}$$

$$SF_x + e^- \xrightarrow{k_{Fx}^*} SF_{x-1} + F + e^- \quad (x = 4,5,6) \tag{5.1.2}$$

$$\begin{aligned} SF_x + O_2 + e^- &\xrightarrow{k_{O_2}} SF_x + O + O^- \\ &\xrightarrow{k_{FOx}} F + SOF_{x-1} + O^- \quad (x = 3,5) \end{aligned} \tag{5.1.3}$$

$$SF_6 + e^- + (\text{other inputs}) \xrightarrow{k_F} F + (\text{other products}) \tag{5.1.4}$$

In the etching processes used in this study as described in the previous section, fluorine is generated by electron-attachment dissociation reaction of SF_6 and many other reactions[5.1.3]. Some of these reactions are described in the reaction equations 5.1.1 - 5.1.3. However, the experimental results discussed before in the fluorine generation model in [5.1.5] indicate that the net effect of all these reactions can be represented by the reaction equation 5.1.4, and the fluorine generation kinetics can be described by a simple generation term $k_F n_e n_{SF_6}$, which is included in Equation 5.1.5 together with other diffusion and loss terms. This is the product of the fluorine generation reaction coefficient (k_F), effective electron concentration (n_e), and SF_6 concentration (n_{SF_6}).

Continuity Equation of (F) Fluorine Concentration n_F :

$$\frac{\partial n_F}{\partial t} = D_F \nabla^2 n_F + k_F n_e n_{SF_6} - \frac{n_F}{\tau_F} \ , \tag{5.1.5}$$

with the boundary conditions $\left.\frac{\partial n_F}{\partial y}\right|_{y=d} = 0$ and

$$\left.\frac{\partial n_F}{\partial y}\right|_{y=0} = \frac{k_{F-Si} n_F}{D_F} \quad \text{for } x < 0 \ ,$$

$$\left.\frac{\partial n_F}{\partial y}\right|_{y=0} = \frac{k_{F-PR} n_F}{D_F} \quad \text{for } x \geq 0 \ .$$

At the top electrode ($y = d$), a zero-flux condition is used as the boundary condition of Equation 5.1.5. At the same time, the boundary condition at the surface of the substrate ($y = 0$) accounts for the fluorine loss due to the surface reactions on the silicon ($x < 0$) and photoresist ($x \geq 0$) surfaces during the etching processes as in the following reaction equations.

Silicon Surface Reaction Equation :

$$Si + 4F \xrightarrow{k_{F-Si}, k_{SiF_4-Si}} SiF_4 \uparrow \tag{5.1.6}$$

Photoresist Surface Reaction Equation :

$$C + 2F \xrightarrow{k_{F-PR}, k_{CF_2-PR}} CF_2 \uparrow \tag{5.1.7}$$

Together with other by-products, SiF_4 and CF_2 are generated from the surfaces of the silicon and photoresist, respectively, according to the reaction equations 5.1.6 and 5.1.7 and the boundary conditions of Equations 5.1.8 and 5.1.9 at the surface of the substrate. At the top electrode ($y = d$), a zero-flux condition is used as the boundary condition as in Equation 5.1.5.

Continuity Equation of SiF_4 Concentration n_{SiF_4} :

$$\frac{\partial n_{SiF_4}}{\partial t} = D_{SiF_4}\nabla^2 n_{SiF_4} - \frac{n_{SiF_4}}{\tau_{SiF_4}} \ , \tag{5.1.8}$$

with the boundary conditions $\left.\frac{\partial n_{SiF_4}}{\partial y}\right|_{y=d} = 0$ and

$$\left.\frac{\partial n_{SiF_4}}{\partial y}\right|_{y=0} = -\frac{k_{SiF_4-Si}n_F}{D_{SiF_4}} \quad \text{for } x < 0$$

$$\left.\frac{\partial n_{SiF_4}}{\partial y}\right|_{y=0} = -\frac{k_{SiF_4-PR}n_F}{D_{SiF_4}} \quad \text{for } x \geq 0 \, .$$

Continuity Equation of CF_2 Concentration n_{CF_2} :

$$\frac{\partial n_{CF_2}}{\partial t} = D_{CF_2}\nabla^2 n_{CF_2} - \frac{n_{CF_2}}{\tau_{CF_2}} \ , \tag{5.1.9}$$

with the boundary conditions $\left.\frac{\partial n_{CF_2}}{\partial y}\right|_{y=d} = 0$ and

$$\left.\frac{\partial n_{CF_2}}{\partial y}\right|_{y=0} = -\frac{k_{CF_2-Si}n_F}{D_{CF_2}} \quad \text{for } x < 0$$

$$\left.\frac{\partial n_{CF_2}}{\partial y}\right|_{y=0} = -\frac{k_{CF_2-PR}n_F}{D_{CF_2}} \quad \text{for } x \geq 0 \, .$$

The diffusion effects of each species are addressed by the first right-hand side terms in Equations 5.1.5 - 5.1.9. Flow effects and radial diffusion are neglected, since neither should affect the concentration profiles greatly within the cross-sectional area of this study. The last terms in Equations 5.1.5 - 5.1.9 account for the bulk loss due to the recombination reactions and the pumping effect with an effective lifetime ($\tau_{F_{Bulk}}$, τ_{SiF_4}, or τ_{CF_2}) for the corresponding species.

Additional assumptions which are needed for the simplified kinetic model are the following. All the parameters, such as the diffusion coefficients (D_F, D_{SiF_4}, D_{CF_2}), reaction constants (k_F, k_{F-Si}, k_{F-PR}, k_{SiF_4-Si}, k_{SiF_4-PR}, k_{CF_2-Si}, k_{CF_2-PR}), and lifetimes (τ_F, τ_{SiF_4}, and τ_{CF_2}), are assumed uniform throughout the reaction volume. Finally, we assume that the concentration of SF_6 (n_{SF_6}) is constant, since only a small fraction of the incoming SF_6 is involved in any reaction for the experimental conditions considered in this chapter.

5.1.4 Computer Simulation

Computer implementation of the dry etching model of the previous section was done with the SUPRA (Stanford University PRocess Analysis) program. This program was originally developed at Stanford to study two-dimensional impurity diffusion in semiconductors, but was extended to treat dry etch processes for this work.

Using parameters from the literature[5.1.6] along with the measured etch rate of other report[5.1.5], the experimental results yield $\tau_F \approx 0.5$ msec. From the process conditions of this study[5.1.5] and the relation[5.1.6] $D = v/3n_o\sigma$, where $v = 6.33 \times 10^4$ cm/sec at 362°K , we can get $D_F = 1.3 \times 10^4$ cm^2/sec, $D_{SiF_4} = 8.6 \times 10^2$ cm^2/sec, and $D_{CF_2} = 2.1 \times 10^3$ cm^2/sec. Fixing these parameters and using the surface reaction coefficient as a single adjustable parameter per species, the simulated steady-state results closely fit the actual data (○ marks) at 2 mm above the wafer surface as in Figures 5.1.2 – 5.1.4. The discrepancy in Figures 5.1.3 and 5.1.4 is due to the measurement error and the other reactions involved in addition to the simple dry etch model. From the computer simulation and fitting procedure, the surface reaction coefficients are extracted to be $k_{F-Si} = 1.04 \times 10^4$ cm/sec, $k_{F-FR} = 3.8 \times 10^3$ cm/sec, $k_{SiF_4-Si} = 3.1 \times 10^8$ cm/sec, and $k_{CF_2-PR} = 2.1 \times 10^7$ cm/sec. Figures 5.1.5 – 5.1.7 are the equi-concentration contour maps of the distribution of each species between the top and bottom electrodes. Finally, Figure 5.1.8 is the three-dimensional plot of the two-dimensional simulation of the distribution of the fluorine atom in the plasma reactor between the two electrodes. It clearly shows the depletion of the fluorine concentration (local loading effect) at the silicon side ($x \leq 0$) due to the fluorine consumption during the silicon etching surface reactions.

This computer simulation is a powerful tool for executing the dry etch model in previous section (Equations 5.1.5 – 5.1.9) into a visualized format as in the equi-concentration contour maps (Figures 5.1.5 – 5.1.7) and three-dimensional plot (Figure 5.1.8). Furthermore, this enables us not only to test out the model by comparing the simulation results to the measured results as in Figures 5.1.2 – 5.1.4, but also to extract the basic plasma parameters through the simulation and fitting procedure as mentioned above for the cases of the surface reaction coefficients. Due to the high diffusion coefficients of the species in the gas state compare to the diffusion mechanisms in the solid state, the simulation predicts the spacial distribution variances of the species concentrations (local loading effects) in the distance of millimeter order instead of micrometer order. These local loading effects are playing important roles in the process parameters such as the etch uniformities across wafers.

5.1.5 Summary

The concentration distribution of fluorine was measured[5.1.5] using Ar as a tracer gas by relating the optical emission associated with fluorine species to its concentration at 2 mm above the wafer surface in the glow discharge. A spectrometer system was designed[5.1.5] to sample the emission from a small sampling volume in order to obtain the highest spatial resolution. The species distributions were modeled using a continuity equation approach which takes into account bulk generation and loss, surface generation and loss, and transport mechanisms. The two-dimensional solution of these equations was carried out using SUPRA. Using parameters extracted from the literature along with measured etch rates, process conditions, and a single adjustable parameter per species, the simulated results closely fit the actual data. This model of the active species, when combined with the surface reaction model of ion bombardment, should yield a powerful model for dry etching

processes.

5.1.6 References

[5.1.1] K. Köhler, J. Coburn, D. Horne, and E. Kay, "Plasma Potentials of 13.56-MHz RF Argon Glow Discharges in a Planar System," *J. Appl. Phys.* 57, 59, 1985.

[5.1.2] D. Flamm, V. Donnelly, and J. Mucha, "The Reaction of Fluorine Atoms with Silicon," *J. Appl. Phys.* 52, 3633, 1981.

[5.1.3] R. d'Agostino and D. Flamm, "Plasma Etching of Si and SiO_2 in SF_6 - O_2 Mixtures," *J. Appl. Phys.* 52, 162, 1981.

[5.1.4] J. Coburn and M. Chen, "Optical Emission Spectroscopy of Reactive Plasma: A Method for Correlating Emission Intensities to Reactive Particle Density," *J. Appl. Phys.* 51, 3134, 1980.

[5.1.5] K. Uhm, M. Kump, J. McVittie, and R. Dutton, "Kinetic Modeling and Measurement of Active Species Distributions during Dry Etching," Proceedings of the Materials Research Soc. Plasma Processing Symposium, Vol. 68, Paper No. C5.3, 1986 MRS Spring Meeting, Palo Alto, CA, April 15 - 19, 1986.

[5.1.6] B. Chapman GLOW DISCHARGE PROCESSES, John Wiley & Sons, NY, 1980.

5.1.7 Figures

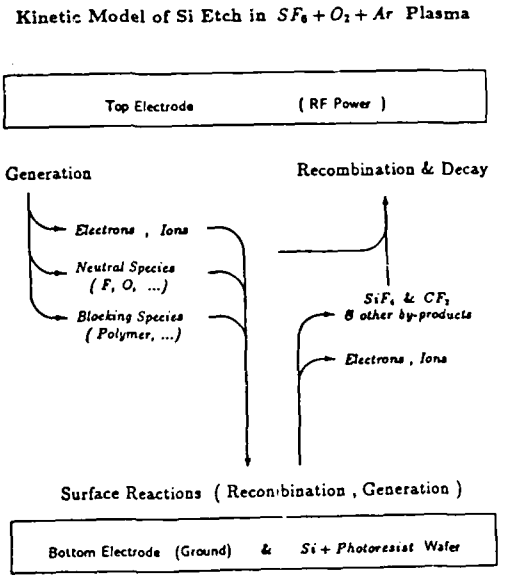


Figure 5.1.1: Schematic of the kinetic model of dry etch system.

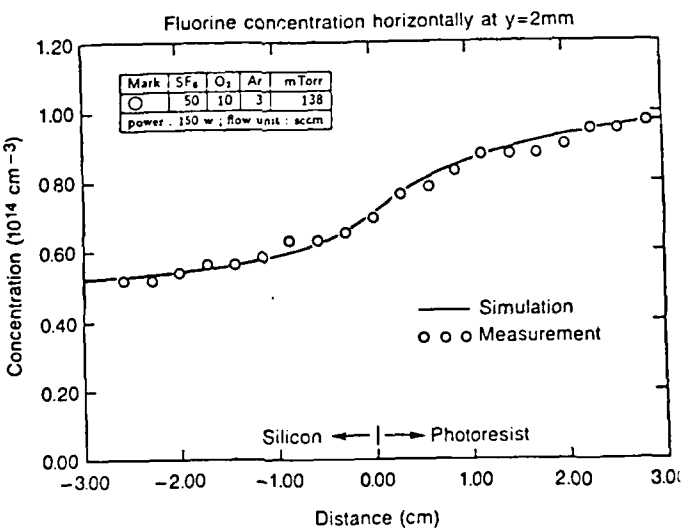


Figure 5.1.2: Comparison between the simulated and measured fluorine concentration distributions at y=2mm.

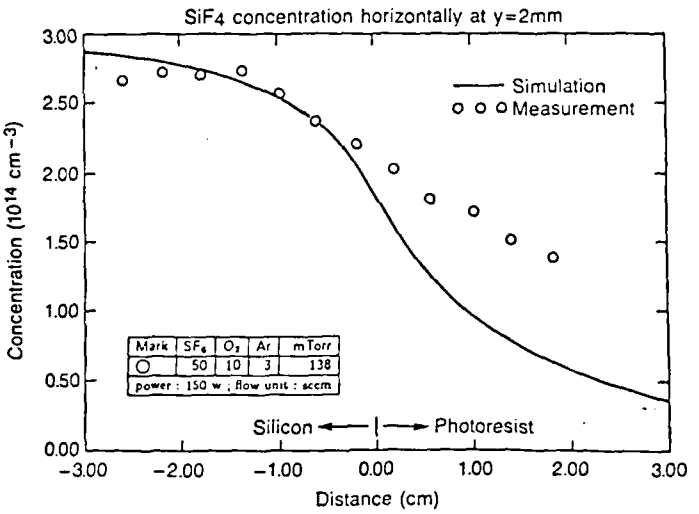


Figure 5.1.3: Comparison between the simulated and measured SiF_4 concentration distributions at y=2mm. The measured data points are in arbitrary unit since they are the ratio of SiF_4 and argon emission intensities.

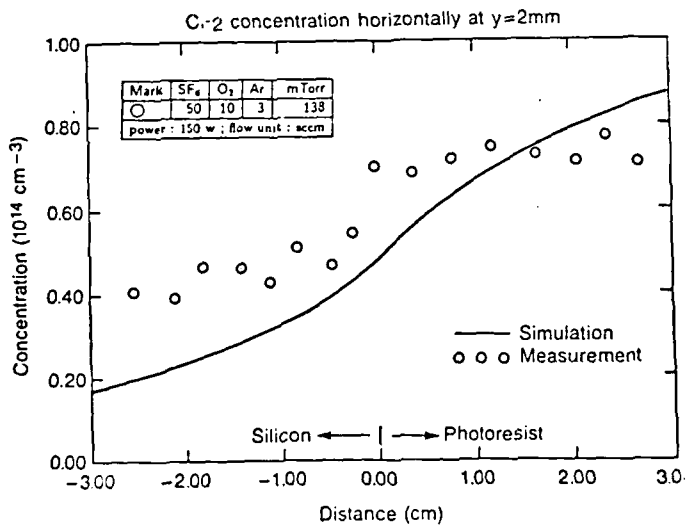


Figure 5.1.4: Comparison between the simulated and measured CF_2 concentration distributions at y=2mm. The measured data points are in arbitrary unit since they are the ratio of CF_2 and argon emission intensities.

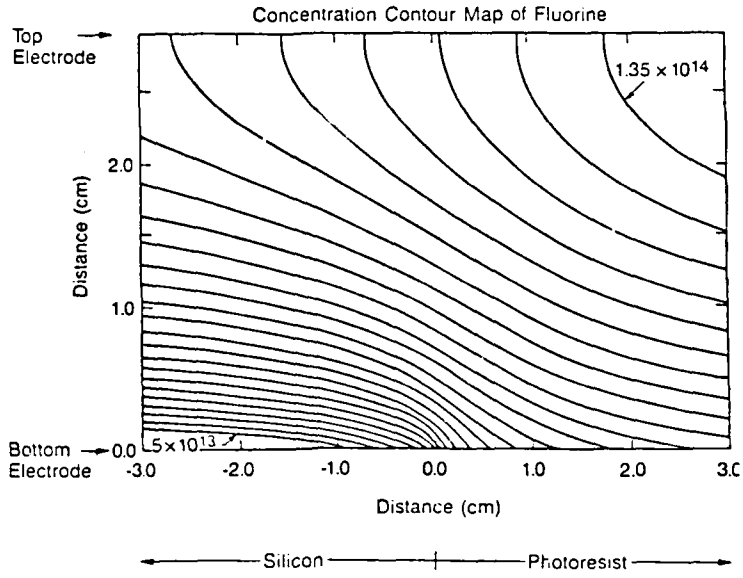


Figure 5.1.5: Equi-concentration contour map of the fluorine distribution from the result of two-dimensional plasma simulation.

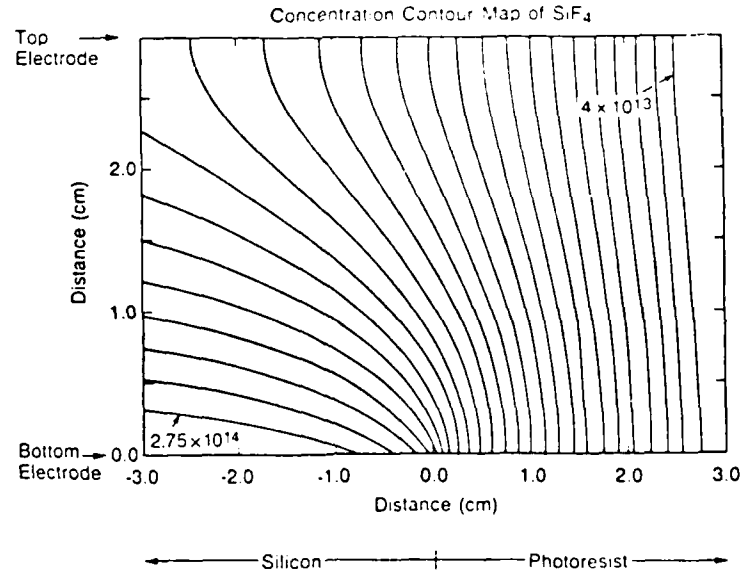


Figure 5.1.6: Equi-concentration contour map of SiF_4 species distribution from the result of two-dimensional plasma simulation.

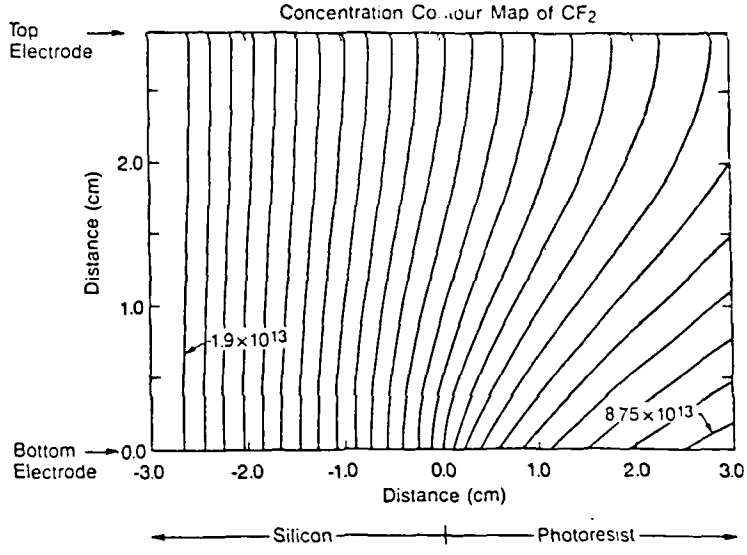


Figure 5.1.7: Equi-concentration contour map of CF_2 species distribution from the result of two-dimensional plasma simulation.

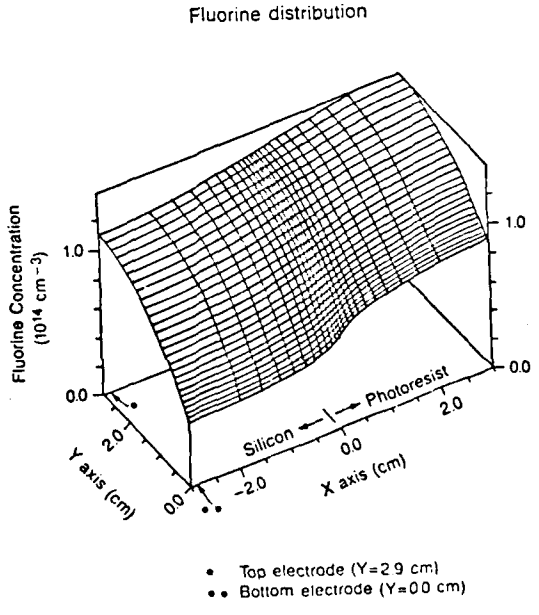


Figure 5.1.8: 3D plot of the SUPRA simulation of fluorine distribution between the top and bottom electrodes.

5.2 In-Situ Monitoring of Electrical Parameters for Dry Etching

J. Ignacio Ulacia F. and James McVittie

5.2.1 Introduction

The understanding and control of dry etch processes has been held up by the lack of knowledge of the internal plasma parameters. Most etch systems only control power, pressure and flows, and do little to directly monitor the plasma discharge set up above the wafer. To improve this situation we are investigating the use of external voltages and currents to monitor the internal plasma parameters. Three parameters of particular interest are the electron density which controls the generation of the active species, the ion current density which determines ion flux onto the wafer, and the sheath thickness which, with the sheath voltage, controls the energy of the ions striking the wafer.

5.2.2 Model

To explain the electrical behavior in a plasma system, we must analyze the different mechanisms responsible for current transport. Externally, with the naked eye, it is possible to distinguish two regions. The first is the bulk of the plasma where a uniform glow fills all the space of the container, and the second is a dark space near the boundaries called the sheath.

In these kind of discharges, usually called partially ionized plasmas, the main species that exist are neutrals. The neutral reactive radicals contribute up to ten percent, while the charged species are 4 to 5 orders of magnitude lower in concentration. The concentration of electrons is very close to that of the ions throughout the bulk region [5.2.3]. This is because diffusion of charged particles in a gas is very fast. If a potential difference appears somewhere in the discharge, electrons and ions will move to compensate for this deficiency. As a result the potential inside the plasma is nearly constant.

To begin our analysis lets investigate the main mechanisms that contribute to current transport. The plasma frequency of the charged particles in the plasma bulk is given by:

$$\omega_{pe}^2 = \frac{n_e q^2}{m_e \epsilon_o} \tag{5.2.1}$$

where ω is the electron plasma frequency, n_e is the electron density, m_e is the mass of the electron, q is its charge and ϵ_o is the permittivity of free space. For typical etch plasma conditions ($n_e = 10^{10} cm^{-3}$) the plasma frequency for electrons is 900 Mhz; while for ions (CF_3^+) it is 2.52 Mhz . This result suggests that the current transport in the bulk of the plasma is carried mainly by electrons. The ions on the other hand only respond to diffusion gradients due to ambipolar diffusion [5.2.29].

Bulk phenomena

With this notion established, lets propose the following simplified model of electron transport that should hold at relatively high pressures ($P > 100mT$) where fluid model approximations are still

valid. The electrons while moving through the bulk suffer collisions with neutral molecules [5.2.17]. This effect can be introduced into Newton's second law for the system as a drag force proportional to the collision frequency, the mass of the electron and its velocity or

$$m_e \frac{\partial \mathbf{u}_e}{\partial t} = -q\mathbf{E}(t) - \nu_{eo} m_e \mathbf{u}_e \tag{5.2.2}$$

where $\mathbf{u_e}$ is the velocity of the electron, $\mathbf{E(t)}$ is the electric field, and ν_{eo} is the electron-neutral collision frequency. If the electric field in the bulk has an oscillation motion of the form $\mathbf{E(t)} = \mathbf{E_o}\, \mathbf{e^{-jwt}}$, the solution of this equation becomes

$$\mathbf{u}_e = -\frac{q}{m_e}\left[\frac{\nu_{eo} + j\omega}{\nu_{eo}^2 + \omega^2}\right] \mathbf{E}(t) \tag{5.2.3}$$

Comparing this result to the current density (J_{eb}), given by:

$$J_{eb} = qn_e\mathbf{u}_e = [\sigma_b + j\omega\chi_b]\, \mathbf{E}(t) \tag{5.2.4}$$

it is possible to identify the conduction current (σ_b) as the real part and displacement current (χ_b) as the imaginary. For modeling purposes in a first order approximation, this mechanism can be described as a resistor and an inductor connected in parallel. In most cases the displacement current carried by the inductor is negligible compared to the conduction current. Neglecting this inductor we get that the current in the plasma bulk is given by

$$J_{eb} = \frac{q^2 n_e \nu_{eo}}{m_e(\nu_{eo}^2 + \omega^2)}\, \mathbf{E}(t) \approx \frac{q^2 n_e}{m_e \nu_{eo}} \tag{5.2.5}$$

where the expression has been simplified since $\nu_{eo}^2 >> \omega^2$.

Sheath phenomena

In the sheath there are many physical phenomena taking place at the same time. In order tc analyze them, it is important to separate the contribution of the ion current transport and electron current transport.

To begin with, the plasma is coupled through a blocking capacitor. This means that no dc current can flow through the system, and that the ion current must balance the electron current over one cycle. As a result a DC bias develops in the system according to the ratio of relative capacitance of both electrodes [5.2.25]. This is usually called the sheath capacitance approximation. There are some precautions that have to be made while considering the DC bias as a modeling parameter. In particular, it is important to consider the secondary emission from the electrodes. This is most important when the surfaces are different materials [5.2.22] (e.g. during etching conditions).

Electrons Initially, the high mobility of the electrons allows them to reach the surface of the electrodes much faster than the ions. This develops a positive potential in the bulk due to electron loss. At this point, and in order for the plasma to retain its condition of electroneutrality, a high negative potential respect to the bulk (and near the electrode surfaces) develops, such that the electrons are rejected from the walls and are retained in the bulk.

There are two main contributions for electron current transport can be divided in two groups, one for electrons that have energy lower than the sheath potential and the other for the ones that have higher energy. Assuming that the electrons have a Maxwellian distribution (not necessarily true) the usual regime of operation of the sheath would fall in the exponential limited regime (figure 1). This implies that most of the electrons have lower energy than the sheath potential and are confined in the bulk of the plasma most of the time. A significant current of electrons only takes place during the portion of the RF cycle when the plasma potential is small respect to the surface. Under this condition bursts of electron current cross the sheath. The electrons with lower energy, are not be able to reach the surface, and they are responsible for a displacement current in the sheath boundaries. Their behavior can be modeled as a nonlinear voltage dependent capacitor, and the plasma reactance is dominated by this sheath capacitance (C_{sh}) given by:

$$C_{sh} = \frac{\epsilon_o A}{l_{sh}} \tag{5.2.6}$$

where A is the electrode area and l_{sh} is the sheath thickness.

The high energy electrons, on the other hand, reach the surface and are collected by the electrode. Previously, several authors have described this mechanism by a reversed biased diode [5.2.1] [5.2.2] [5.2.4] [5.2.15] [5.2.18] [5.2.17] [5.2.25] but actually it is best modeled as an exponential voltage dependent current source, similar to a diode. The difference is that in a diode there is no net current flow for zero applied voltage, while for an exponential voltage dependent current source, current flows even if the voltage across it is zero.

This exponential voltage dependent current source has the following behavior

$$J_e = \frac{n_e \bar{u}_e}{4} e^{V_{sh}/kT_e} \tag{5.2.7}$$

where J_e is the electron current density, $\bar{u}_e$ is the electron average thermal velocity, k is the Boltzman constant and T_e is the Maxwellian electron temperature.

Ions The ions on the other hand only see a down hill potential that accelerates them to the surface. This is the cause of the highly energetic ion bombardment in plasma etching or Reactive Ion Etching (RIE). The ion energy at the wafer surface depends on the sheath voltage and the ratio of sheath thickness to the mean free path. The current density is limited by ion mobility and the average dc sheath potential.

The motion of the ions while transversing the sheath is governed by the Child-Langmuir law [5.2.3] that has the form

$$J_i = \frac{4\epsilon_o}{9}\left(\frac{2q}{m_i}\right)^{1/2}\left(\frac{V_{sh}^{3/2}}{l_{sh}^2}\right) \tag{5.2.8}$$

where J_i is the ion current density, and m_i is the ion mass. This behavior can be modeled as a non linear resistor.

An additional source of ion current is secondary electron emission; either by ion impact or by photoemission. In either case this process can be modeled as a current dependent current source. Since the yield for secondary electron emission is relatively low for the energies considered we have ignored this contribution.

Shown in figure 2.a is the electrical circuit model in which only the most important physical current transport mechanisms have been considered. The central plasma region is modeled by a resistor corresponding to electron-neutral collision losses. The sheaths are modeled as an average value resistor in parallel with a capacitor and the 'diode-like' current source. The resistor describes the ion current, the capacitor that accounts for the low energy electrons and the 'diode-like' voltage dependent current source for the high energy electrons.

During the rf excitation, the diode-like current source will 'short circuit' the system at the cycle peaks. In these conditions the circuit reduces to figure 2.b. From circuit theory, the impedance can be calculated as:

$$Z = Z_r + jZ_i = R_b + \frac{Rsh}{1+\omega C_{sh} {R_{sh}}^2} - j\omega \frac{C_{sh} R_{sh}^2}{1+\omega C_{sh} {R_{sh}}^2} \tag{5.2.9}$$

where Z is the total impedance, Z_r and Z_i the real and imaginary impedances, R_b the bulk resistance, R_{sh} the sheath resistance, and C_{sh} the sheath capacitance.

This treatment isolates the different contributions of each mechanism to the total current transport, and allows the calculation of important parameters of the discharge. In table 1 we can find a summary of the electrical circuit components related to the plasma internal parameters.

From the bulk resistance we can calculate the electron energy and the collision frequency by:

$$n_e = \frac{\sigma_b m_e \nu_{eo}}{q^2} \tag{5.2.10}$$

$$\nu_{eo} = n_o Q_{eo} \left(\frac{kT_e}{m_e}\right)^{1/2} \tag{5.2.11}$$

where n_o is the neutral concentration (cm^{-3}), and Q_{eo} is the electron-neutral total collision cross section. The only disadvantage is that an independent measurement or calculation of electron temperature must be available, either from a Langmuir probe measurement or by the solution of the Boltzman equation.

From the sheath capacitor, the sheath thickness by:

$$l_{sh} = \frac{\epsilon_o A}{C_{sh}} = \epsilon_o A \omega Z_i \tag{5.2.12}$$

and from the sheath resistor, the ion current density calculated by the Child-Langmuir law previously described.

5.2.3 Experimental

In order to verify the use of external impedance measurements and to obtain the internal plasma parameters during dry etching, we have studied a $SF_6 : O_2$ discharge in a parallel plate etch system operating in the plasma mode (high pressure and the wafer placed on the grounded electrode). This gas mixture is currently used in many semiconductor manufacturing processing steps, such as polysilicon etching and tungsten silicide [5.2.31] [5.2.21].

Our experimental setup consists of a parallel plate etcher, Plasma-Therm PK-1241. The electrode area and spacing are 510 cm^2 and 2.9 mm respectively. Both electrodes are cooled and are kept at 25 C. The impedance of the discharge is calculated indirectly by the measurement of the voltage, current and phase angle. In figure 3 we see the circuit used. The voltage, dc and ac, are

recorded through a 60 dB attenuated transmission line terminated in a high and low pass filters respectively. The current was obtained from a 20:1 current transformer attenuated by a pi network, 40 dB in total. The system was calibrated and the parasitics characterized before usage, the losses do to parasitic components were almost negligible but were considered in our analysis, and an agreement from power readings to electrode calculated power agreed within 5

In figure 4 we can observe the impedance as a function of the reciprocal of current density for different conditions of power and pressure. It can be seen that the behavior of the discharge is characterized by the model and that the impedance is linear at constant pressure. At low current densities the curves saturate probably due to discharge instabilities. For constant power, a parabolic behavior is observed, same that can be described by the model.

From this data it is possible to extract the electron density, the ion current density and the sheath thickness values that are shown in figures 5, 6 and 7. The value of electron temperature used in the calculation was obtained from a Langmuir single probe experimental data previously obtained in our laboratory $T_e = 6.4eV$, this value has been verified using the solution of the Boltzman transport equation to be $T_e = 6.21eV$ We have assumed that this value holds true for all the regime since the electron temperature has a big dependence on pressure but not on power density. For this reason the pressure data has not been considered. These curves should be the subject of future work.

5.2.4 Conclusions

From the model developed and the data obtained, we have been able to calculate and assign values of electron density, ion current density, sheath thickness and electrical circuit components to the SF_6 . O_2 discharge. With these values the system can be characterized and the model can be fed into a process modeling program.

The advantages obtained in the use of this method is that the technique is non-invasive and can be incorporated into a plasma etcher with a minimum of hardware.

In a future perspective we have several goals:

a) To fully characterize the technique and to apply it to other etching environments such as CF_4, O_2, $SF_6 : C_2ClF_5$ and sputtering environments such as Ar plasmas.

b) To make a sheath ion energy simulator by the Montecarlo method to obtain information on the angular distribution and ion energies on the sheath surface.

c) To generate a plasma etching model that includes the contributions due to chemical etching, ion bombardment and inhibitors species created as byproducts in the etching environment.

5.2.5 References

[5.2.1] H.R.Koenig and L.I. Maissel, "Application of RF discharges to sputtering", IBM J. Res. Develop. 14,[3],168,(1970).

[5.2.2] J.W. Couburn and E.Kay , "Positive-ion bombardment of substrate in rf glow discharge sputtering", J. Appl. Phys 43,[12],1965,(1972)

[5.2.3] F.F. Chen , "Introduction to Plasma Physics and Controlled Fusion", Plenum Press, Second edition, New York (1984).

[5.2.4] J.S. Logan, J.H. Keller and R.G. Simmons, "The rf glow discharge sputtering model", J. Vac. Sci. Technol.,**14**,[1],92,(1977)

[5.2.5] J.H. Keller and W.B. Pennebaker ; "Electrical properties of RF sputtering systems" ; IBM J. Res. Develop. **23**,[1],3,(1979)

[5.2.6] H. Norstrom ; "Experimental and design information for calculating impedance matching networks for use in rf sputtering and plasma chemistry" ; Vacuum **29**,[10],341,(1979)

[5.2.7] W.B. Pennebaker ; " Influence of scattering and ionization on RF Impedance in glow discharge sheaths " ; IBM J. Res. Develop.

[5.2.8] K. Ukai and K. Hanazawa ; "End-point determination of aluminum reactive ion etching by discharge impedance monitoring" ; J. Vac. Sci. Technol. **16**,[2],385,(1979)

[5.2.9] A.R. Tretola ; "Method of controlling a plasma etching process by monitoring the impedance changes of the RF power" ; US patent 4,207,137

[5.2.10] D.C. Ilic ; "Impedance measurement as a diagnostic for plasma reactors"; Rev. Sci. Instrum. **52**,[10],1542,(1981)

[5.2.11] R.A. Morgan ; "The rf voltage/current characteristics and related dc negative bias properties of an Electrotech flat bed plasma etcher" ; Vacuum **32**,[5],297,(1982)

[5.2.12] C.B. Zarowin and R.S. Horwath ; "Control of plasma etch profiles with plasma sheath electric field and RF power density" ; J. Electrochem. Soc. **129**,[11],2541,(1982)

[5.2.13] K.P. Brad and H. Jungblut ; "The interaction potentials of SF6 ions in SF6 parent gas determined from mobility data" ; J. Chem. Phys. **78**,[4],1999,(1983)

[5.2.14] S. Broydo ; "Important considerations in selecting anisotropic plasma etching equipment"; Solid State Technol. [4],159,(1983)

[5.2.15] C.M. Horwitz, "Rf sputtering-voltage division between two electrodes", J. Vac. Sci. Technol.,A **1**,[1],60,(1983)

[5.2.16] Mantei, T.D.; J. Electrochem. Soc. **130**,[9],1958,(1983)

[5.2.17] W.E. Mlynko ; "Electrical diagnostics of radio frequency glow discharges used for plasma etching" ; Chemical Engineering, UC-Berkeley (1983)

[5.2.18] J.L. Reynolds ; "Characterization of plasma etched structures in IC processing" ; Electrical Engineering and Computer Science, UC-Berkeley (1983).

[5.2.19] A.J. Van R[illegible]en ; "Plasma parameter estimation from rf impedance measurements in a dry etchin[illegible]em" , Appl. Phys. Lttr. **42**,[5],416,(1983)

[5.2.20] C.B. Zarowin ; "Plasma etch anisotropy - Theory and some verifying experiments relating ion transport, ion energy and etch profiles"; J. Electrochem. Soc. **130**,[5],1144,(1983)

[5.2.21] S.E. Clark, J. K. Tsang, and J.W. Marolf , "Deposition and patterning of Tungsten and Tantalum polycides", Solid State Technol. [4],235,(1984)

[5.2.22] M.D. Gill ; "Sustaining mechanisms in rf plasmas" ; Vacuum **34**,[3-4],357,(1984)

[5.2.23] J.P. Novak and M.F. Frechette ; "Transport coefficients of SF6 and SF6-N2 mixtures from revised data"; J. Appl. Phys. **55**,[1],107,(1984)

[5.2.24] C.B. Zarowin ; "Relation between the RF discharge parameters and plasma etch rates, selectivity, and anisotropy" J. Vac. Sci. Technol. **A2**,[2],1537,(1984)

[5.2.25] K. Kohler, J.W. Coburn, D.E. Horne, and E. Kay., "Plasma potential of 13.56-MHz rf argon glow discharges in a planar system" ; J. Appl. Phys. **57**,[1],59,(1985)

[5.2.26] K. Kohler, D.E. Horne, and J.W. Coburn, "Frequency dependance of ion bombardment of grounded surfaces in rf argon glow discharges in planar systems", J. Appl. Phys. **58**,[9],3350,(1985)

[5.2.27] O.A. Popov and V.A. Godyak ; "Power dissipated in low-pressure radio-frequency discharge plasmas" ; J. Appl. Phys. **57**,[1],53,(1985)

[5.2.28] A.J. Van Roosmalen et al.; "Electrical properties of planar rf discharges for dry etching" ; J. Appl. Phys. **58**,[2],653,(1985)

[5.2.29] D.B. Graves ; "A continuum model of low pressure gas discharges" ; University of Minnesota (1986)

[5.2.30] G.A. Hebner and J.T. Verdeyen ; "The spatial and temporal evolution of the glow in an RF discharge" ; IEEE Transactions on plasma science **PS-14**,[2],132,(1986)

[5.2.31] J. Herrmann, J. Oncay, and R. Khathuria, "Anisotropic plasma etching of polysilicon using SF_6/O_2 in a high pressure plasma etcher", Lam Research Corporation, Fremont, California. 490-TA-0386

[5.2.32] A. Paranajpe, and S.P. Kerckhoff , private communication.

[5.2.33] W.G.M. van den Hoek , et. al. ; "Power loss mechanisms in radio frequency dry etching systems" ; MRS symposia march (1986)

5.2.6 Figures

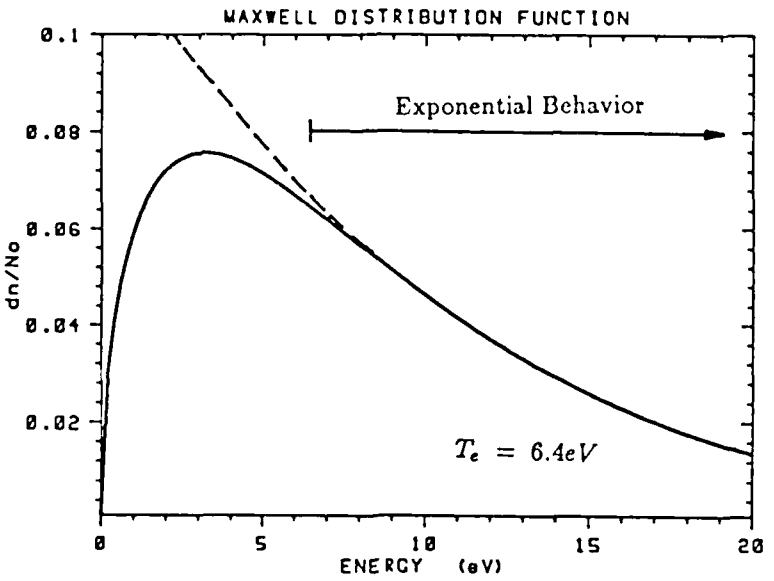


Figure 5.2.1: Maxwell-Boltzman energy distribution function for an electron temperature of 6.4 eV. Note that at energies greater than the electron temperature, the curve is limited by the exponential.

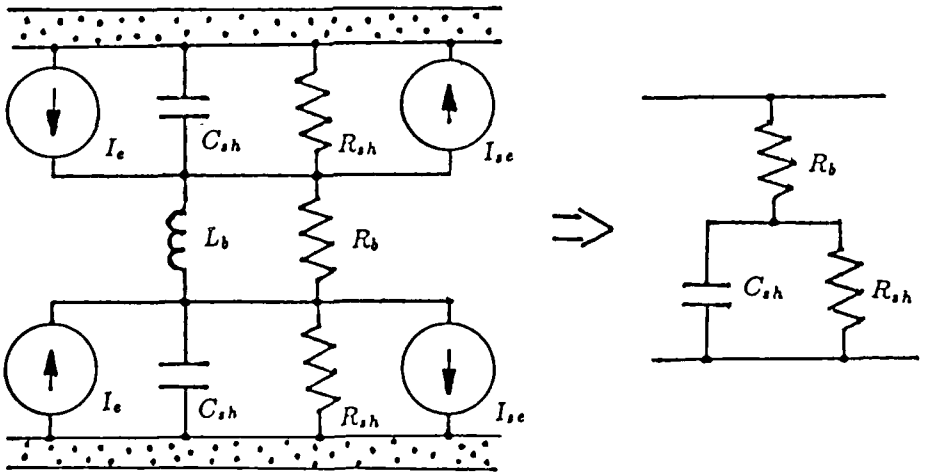
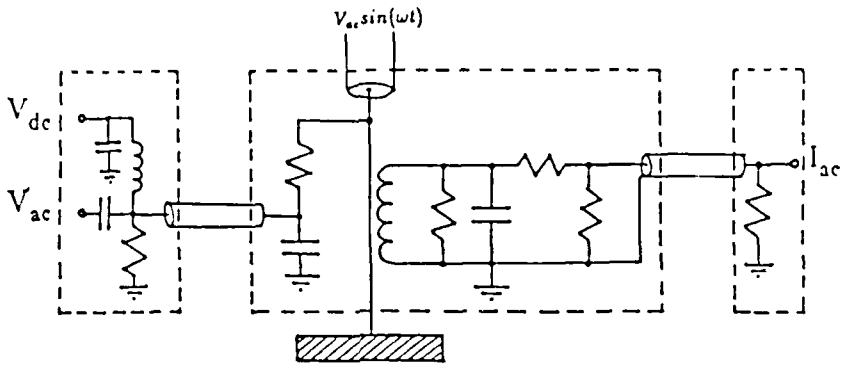


Figure 5.2.2: Schematic of the plasma circuit diagram. The bulk resistor (R_b) represents the loss due to collisions. The sheath resistors (R_{sh}) represent the ion current in the sheath. The current source (I_{se}) secondary emission due to ions bombardment and UV radiation. The capacitor (C_{sh}) and the current source (I_e) represent the low and high energy electrons respectively. In the second form represents the time average measured circuit.



$$Z_p = \frac{V_{ac}sin(\omega t)}{I_{ac}sin(\omega t + \theta)}$$

Figure 5.2.3: Measurement circuit. Consists of a voltage divider and a current transformer attenuated and filtered.

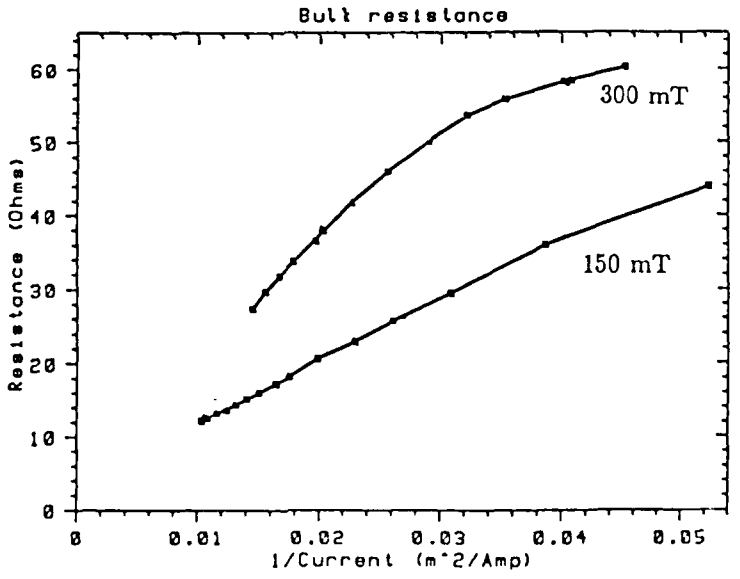


Figure 5.2.4: Bulk resistance as a function of the reciprocal of current density. Observe the linear behavior at constant pressure, saturating at low values of current density, perhaps due to plasma instability.

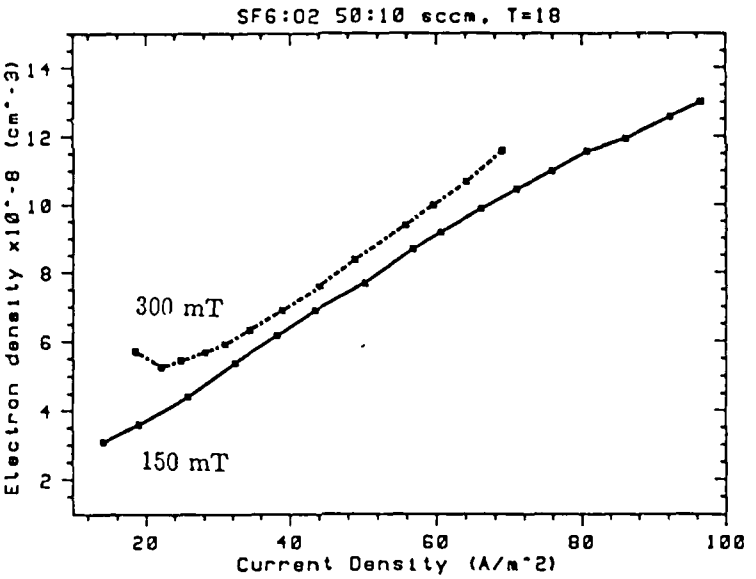


Figure 5.2.5: Electron density calculated from the model. These results agree with the solutions to the Boltzman transport equation.

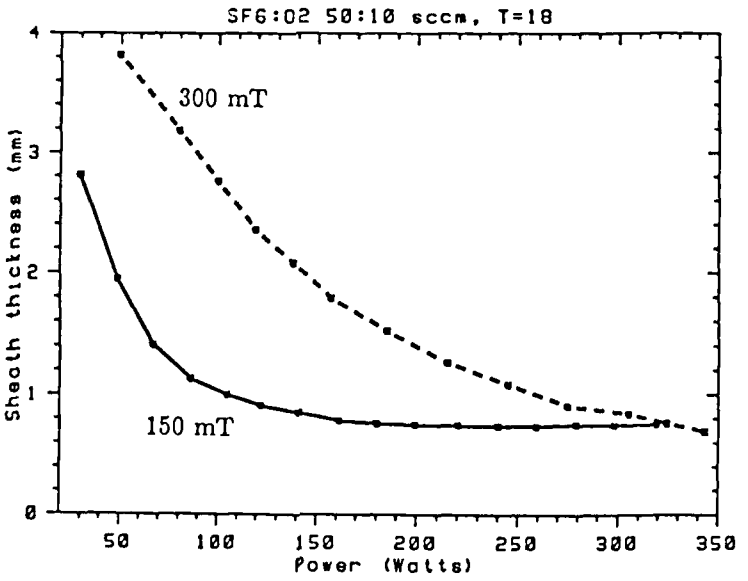


Figure 5.2.6: Calculated sheath thickness, from the model.

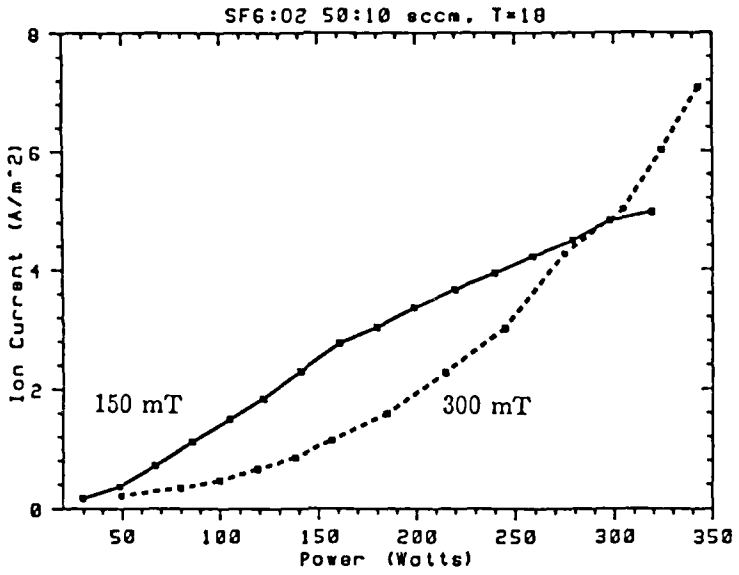


Figure 5.2.7: Calculated Ion current density, from Child-Langmuir law.

Electron density	$n_e = \frac{m_e \nu_{eo} l_{el}}{q^2 R_b A_{el}}$
Sheath thickness	$l_{sh} = \epsilon_o \omega Z_i A_{el}$
Ion current flux	$J_i = \frac{4\epsilon_o}{9}\left(\frac{2q}{m_i}\right)^{1/2}\left(\frac{V_{sh}^{3/2}}{l_{sh}}\right)$

Table 5.2.1: Formulas used for the calculation of electron density, Sheath thickness, and Ion current flux, knowing the impedance and plasma potential of the discharge.

5.3 Sidewall Residues in Dry Etching

A. Bariya and J.P. McVittie

5.3.1 Introduction

During plasma etching, several different sub-processes occur simultaneously and in an interactive manner. These sub-processes are:

- Chemical etching.
- Ion bombardment of the surfaces exposed to the discharge.
- Residue formation on the substrate.

To better understand the overall process, these sub-processes need to be separated and individually studied. Understanding their individual effects and their modes of interaction will lead to realistic models of plasma etching.

Residue formation on the substrate during plasma etching is an important issue in optimizing the process. Residues form as a result of polymerisation reactions occurring between the free radicals created in the discharge. They may be a nuisance, affecting subsequent processing steps such as deposition and requiring additional steps for their removal. Further, they may reduce the etch rate. However, they can play an important role in obtaining anisotropic etch profiles; especially in single wafer etchers where higher pressures and lower ion energies are used. In these cases anisotropy is a result of the interaction between residue formation and ion bombardment. Residues form on the sidewalls as well as the floor of the etch profile. Ion bombardment, however, is restricted only to the floor since ions are incident normally onto the surface. This results in thinner residue layers forming on the floor as compared to the sidewalls. The etch rate on the sidewalls is therefore lower than on the floor, due to the greater inhibiting effect of the residues on the former. The result: an anisotropic etch profile. Residues also help in getting better selectivities. For instance, during oxide etching, the presence of oxygen in the oxide decreases the concentration of the residue forming precursors. Once the oxide is etched through and the underlying Si exposed, there is no longer a source of oxygen. Residues form rapidly and the etch rate of Si plummets, leading to higher oxide to Si selectivity. Thus we see that residues have both a beneficial and deleterious effect.

An effort is underway to study the nature of these residues and the effect of ion bombardment on them. This is accomplished by stopping, or diminishing the energy and flux of, the impinging ions. This enables the study of the effect of ion bombardment on the nature of the residue and at the same time, allows sufficient residue to be accumulated for surface analysis to be possible. Ion bombardment is reduced by what we refer to as the "grid" technique. The rest of this report will deal with a description of this technique and some of the preliminary results obtained.

5.3.2 Grid Technique

The grid technique is schematically illustrated in fig.1. It involves placing the substrate, which lies on the grounded electrode, underneath a grounded aluminum grid. There is no field between the

grid and the electrode since they are both at the same potential. Ions are accelerated across the sheath between the discharge and the grid, but once they penetrate the grid, there is no field to accelerate them further. The distance between the grid and the substrate is many mean-free-paths (of the ion) and hence the ion suffers several collisions before it reaches the substrate. This causes it to lose the energy it had gained during its transit through the sheath, and if it eventually hits the surface, it does so with only its thermal energy.

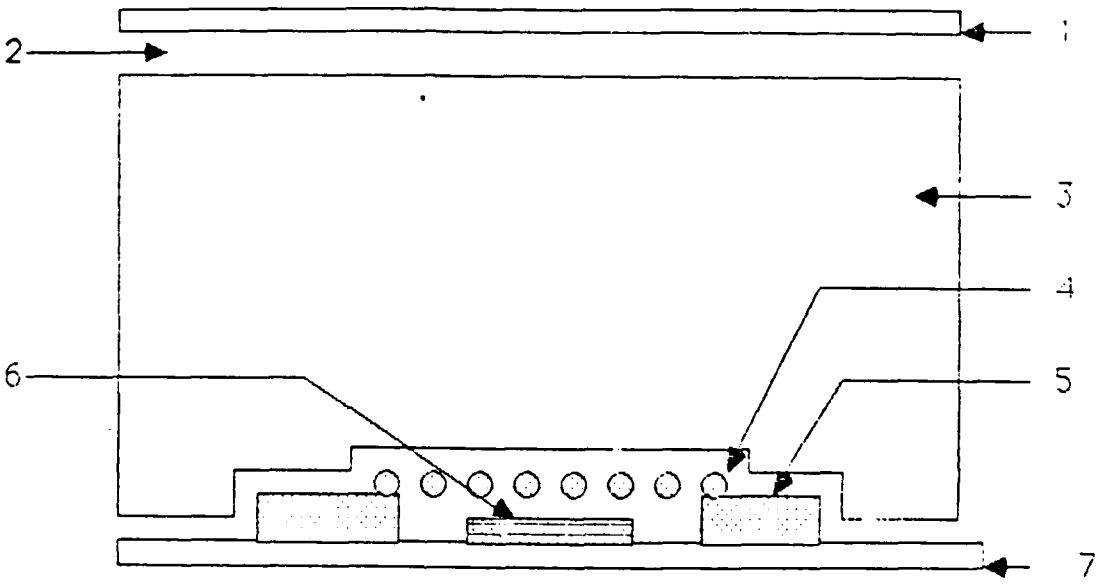
5.3.3 Results

Residue was collected on a piece of bare Si wafer placed under the grid as well as on a control piece over which there was no grid present. The etch conditions were:

Pressure	150 mTorr
Power	500 watts
Gas	Flow Rate
Sulfur hexafluoride	150 sccm
Freon 115	150 sccm
Etch time	20 minutes

There were two sets of samples. One set involved the presence of AZ1470 positive photoresist on the aluminium plate supporting the grid. For the other set, no photoresist was present. As soon as the etching was complete, the samples were put into an X-ray photoelectron spectrometer for surface analysis. Time of exposure to the ambient was minimized to the extent possible. The XPS results are shown in fig.2. A large graphitic carbon peak is shown to be present for all cases except the resist + grid case where it is smaller than some of the other, "bonded", peaks. It appears from the spectra that the ion bombardment is affecting the nature of the residue formed.

Prior to this experiment, a similar experiment was performed in which no surface analysis was done but instead the etched bare Si wafer pieces were surface-profiled on a dektak. The results of this surface profiling are presented in fig.3. The effectiveness of the grid technique is apparent in these results. Notice the deeper etch in the sample placed outside the grid (open) than the one placed under it. Also notice the drastically lower etch for the resist case when the sample was under the grid. This demonstrates that the resist is a source of residue forming precursors and that the ion bombardment is effective in reducing this residue.



1 Top (powered) electrode
2 Sheath
3 Discharge
4 Grounded aluminium grid
5 Aluminium support for the grid (grounded)
6 Silicon substrate
7 Bottom (grounded) electrode

Figure 5.3.1: Grid Technique

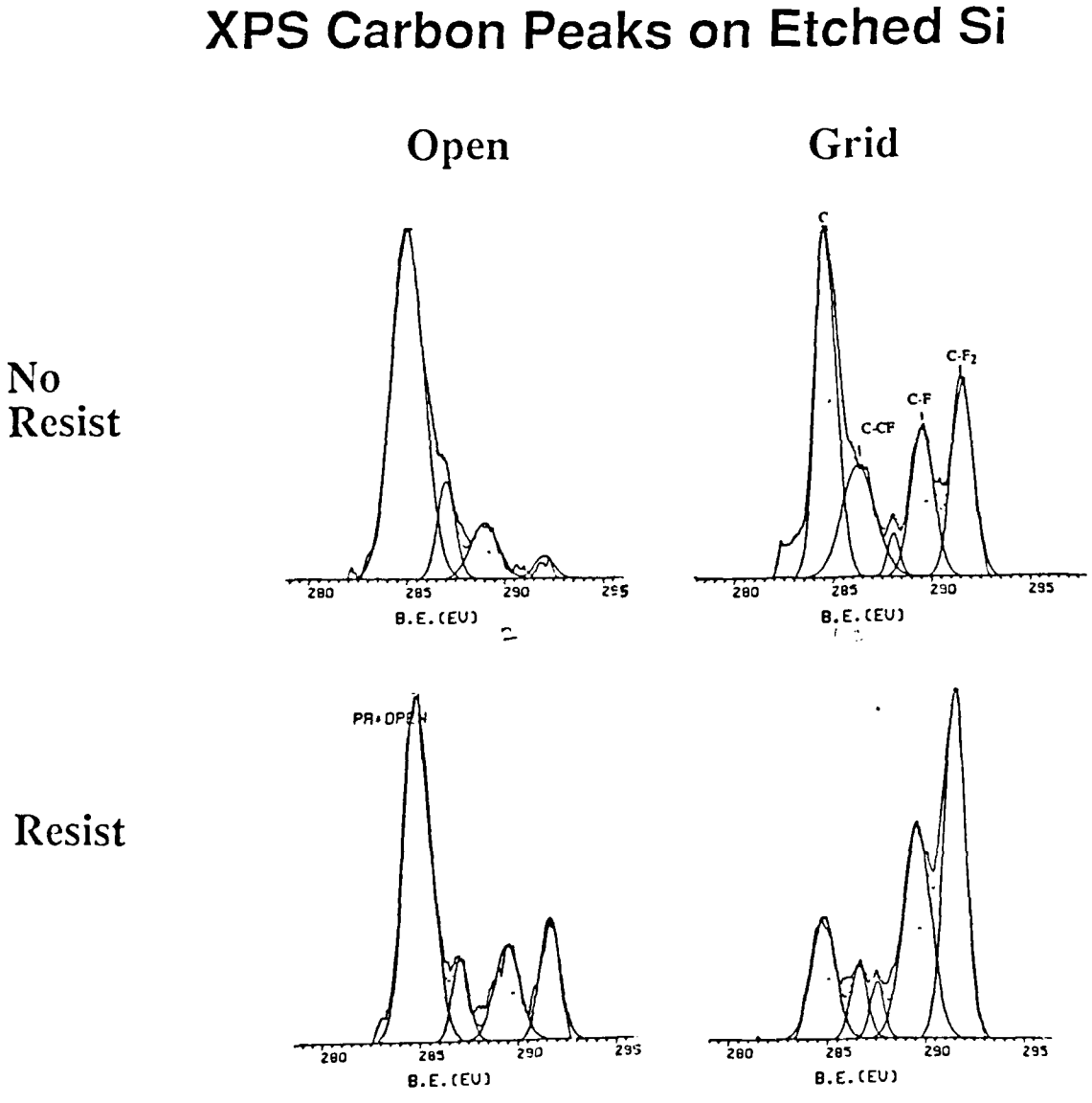
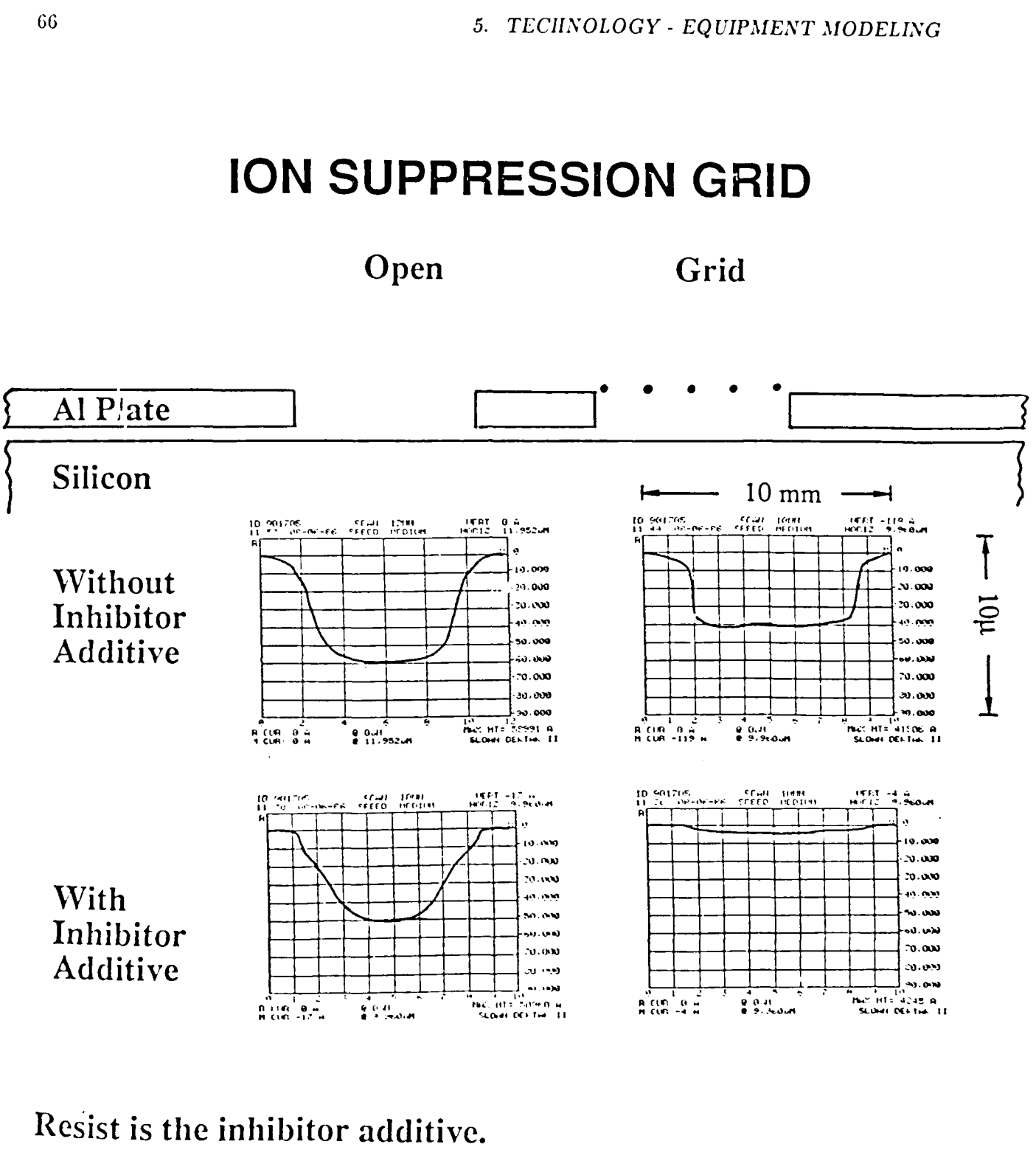


Figure 5.3.2: XPS Carbon Peaks on Etched Si

ION SUPPRESSION GRID



Resist is the inhibitor additive.

Figure 5.3.3: Ion Suppression Grid

5.4 Rapid Thermal Processing for Advanced MOS Technologies

Mehrdad M. Moslehi and Krishna C. Saraswat

5.4.1 Introduction

Rapid thermal processing (RTP) is an emerging technology with many important applications in silicon integrated circuits and compound semiconductors [5.4.1]-[5.4.7]. In the past few years, we have been investigating some novel applications of RTP including thin dielectric growth by rapid thermal oxidation and nitridation processes [5.4.8]-[5.4.15] and reactive-ambient annealing of refractory metals [5.4.15]. The results have been very promising and we will continue these studies with the goal of placing further emphasis on manufacturability of the RTP applications. The main objectives of this work are to identify and analyze the RTP equipment parameters and manufacturing requirements through its conventional and novel applications and by employing an advanced RTP-based submicron CMOS technology as a technology vehicle. Because RTP is a multipurpose technique (annealing, growth, and deposition processes), the equipment and process models are application-dependent. As a result, a comprehensive physical understanding of these process applications is necessary for developing appropriate application-specific equipment models.

This report presents a summary of the fundamental properties of the RTP equipment and techniques and our recent research results on the rapid thermal growth of thin gate dielectrics. As part of these studies [5.4.16], metal-oxide-semiconductor devices fabricated with tungsten/n^+ polysilicon composite gates and subhundred-angstrom SiO_2 gate insulators grown by rapid thermal oxidation were characterized by various electrical measurements. The as-fabricated devices with unannealed rapidly grown oxides exhibited breakdown characteristics superior to furnace-grown oxides as evidenced by their excellent breakdown uniformity, an average breakdown field of 15 MV/cm, and an average breakdown charge density of over 50 C/cm^2 at a stress current density of 1 A/cm^2. The preoxidation surface cleaning procedure was observed to affect the charge-to-breakdown and the densities of fixed oxide charges and surface states in these MOS structures.

5.4.2 Rapid Thermal Processing of Silicon Wafers

Following its introduction for activation of implanted dopants [5.4.2], RTP has proved to be useful in a variety of processing applications. There are many important requirements in advanced silicon technologies which deserve careful considerations. One such need in scaled MOS processes is the formation of very shallow (source/drain) junctions with the implication that the dopant redistribution caused by thermal processing must be minimized. A limited thermal budget rules out the use of any long high-temperature processing steps. However, some processes are in favor of certain high-temperature cycles to achieve better electrical performance in fabricated devices. Examples include ion implant activation and damage removal, resistivity reduction of silicides, glass reflow, and gate dielectric growth. RTP offers a compromise because it allows the higher temperature processing but for very short times (a few seconds). Fortunately, the extent of undesirable processes such as dopant redistribution is significantly reduced in the short-time regime. The rates of many processes such as surface nitridation of SiO_2 or implant activation are strongly activated with temperature and, therefore, they proceed rapidly during the very early stages of the high-temperature

anneal cycles. In contrast to furnace processing, RTP can selectively drive the desirable processes and suppress the unwanted ones such as diffusion (broadening).

In general, the wafer heating techniques are classified as

- resistively heated furnaces (conventional approach)
- CW or pulsed scanned beams of energy (laser or electron beams)
- rapid isothermal bulk heating (incoherent light, blackbody source, microwave, charged beams).

The rapid heating technique employs a directed energy beam to heat the surface or bulk of a wafer for time periods ranging from several nanoseconds up to several minutes depending on the type of the energy source. The heating time period for any spot on a wafer depends on the heating technique. The pulsed and CW scanned beams only heat the surface or near-surface region leaving the bulk of the wafer at a lower temperature. Some processes can take advantage of selective heating of the surface region. One example is the phosphosilicate glass reflow over aluminum by a pulsed or CW scanned heating technique. An RTP system with incoherent light source is used for isothermal and uniform wafer heating for a wide range of times from a few seconds to a few hundred seconds.

Wafer heating techniques are usually based on one or a combination of the following physical mechanisms:

- convection (conventional furnaces)
- mechanical momentum transfer (by CW or pulsed ion and electron beams such as in a plasma ambient)
- radiation (optical, microwave).

A major portion of the radiation energy in a tungsten-halogen lamp is in the near-infrared region where the free-carrier heating mechanism dominates. Most of the commercial RTP systems (e.g. AG Associates, AET Addax, and Varian/Extrion [5.4.17]) including the AET ADDAX R 1000 at CIS accomplish optical heating through free-carrier absorption. The strongest absorption in silicon at room temperature occurs at visible and UV ($\lambda \leq 1.1\ \mu$m) region of the electromagnetic spectrum. However, at longer wavelengths and in the infrared region, the free-carrier absorption is the dominant heating mechanism. Free-carrier absorption occurs even with the incident photon energies much less than the silicon bandgap because these low-energy photons can excite the free carriers in the semiconductor band edges which then generate lattice phonons and wafer heating. The free-carrier absorption in unheated wafers increases as the background doping level is raised. At room temperature and in lightly doped wafers, the free-carrier absorption is small because of the low concentration of free carriers (equal to the doping density). As the wafer is gradually heated, additional free carriers will be thermally generated (n_i or the intrinsic carrier concentration is a strong function of temperature) and, as a result, the absorption coefficient increases rapidly with temperature. The free-carrier absorption depends on the doping concentration; however, the effect of background doping is diminished at higher temperatures where silicon is usually in the intrinsic regime and the intrinsic carrier concentration dominates the background doping and wafer electrical conductivity.

5.4.3 RTP Chamber Design

The processing chamber can be designed to act as an absorbing heat sink or to be reflecting or a combination of them. Absorbing chamber walls are useful for rapid wafer cooling during the postprocess ramp-down period and reduce the run-to-run chamber heat memory effects but demand high power consumption during wafer heating. On the other hand, reflecting walls have exactly the opposite effects resulting in poor reproducibility and throughput. Therefore, there is a trade-off between wafer cooling rate during ramp-down (or throughput) and RTP source power consumption. In an open-loop mode of operation, the wafer heating rate depends on doping level and its activation and wafer thickness. Some processes are strongly affected not only by the steady-state heating cycle but also by the temperature ramp-up rate. An example is rapid thermal nitridation of oxidized silicon in which the temperature ramp-up rate in ammonia ambient has an effect on the nitrogen compositional profile of the grown nitroxide dielectrics. As a result, the equipment and process parameters in RTP have strong correlations which should be employed for equipment characterization and modeling purposes. The RTP cooling rate depends on the chamber temperature and may be much slower than the ramp-up cycle in most commercial RTP systems. The RTP system throughput is a function of the process to be performed. Simple inert ambient anneals for dopant activation demand the least strict requirements in terms of wafer cooling and preprocess purge and, as a result, have high throughput. In contrast, a process such as tungsten anneal in inert or ammonia ambients requires a complete preprocess purge after wafer loading and postprocess wafer cooling (to less than 200°C) in a nonoxidizing controlled ambient before unloading the wafer. Removing a hot wafer from its controlled chamber ambient may result in device degradations caused by the effects of oxygen, moisture, and other impurities in the air. Advanced RTP systems will employ water-cooled metallic chamber and input/output loadlocks to enhance manufacturability and process throughput and to reduce the risk of process-induced contamination.

Development of advanced manufacturable integrated circuit processes is not possible unless all sources of process- and equipment-induced contamination are identified and eliminated. Contamination which is known to cause yield degradations originates from a variety of sources. Interaction of exposed metallic parts in the chamber with process gas and the residual impurities in the gas are some examples of possible contamination sources. Generation of particulates must be minimized by proper equipment design and automated wafer handling.

5.4.4 RTP Equipment Parameters

RTP offers a time-temperature process window useful for kinetics-selective processing so that only the desirable process proceeds and the unwanted processes are suppressed. It has proved to be applicable to fast turnaround process development and formation and processing of new materials. It is also an ideal tool for fast wafer processing because the thermal loading is usually much less than that in a furnace. In most of the high-temperature processes the effect of doping on steady-state wafer temperature is negligible. However, its effect on the ramp-up rate is strong because, at lower temperatures, silicon is initially in the extrinsic doping regime and the free-carrier concentration and absorption coefficient are both functions of temperature. Some commercial systems (e.g. Peak Systems [5.4.18]) employ optical sources radiating mostly in the visible and UV range to promote the intrinsic absorption in silicon. Because the intrinsic absorption of silicon is independent of the background doping and is strong irrespective of the wafer temperature, more reproducible and

faster temperature ramp-ups can be achieved with these energy sources. On the other hand, one should be aware of the possibility of gas dissociations and formation of radicals and even wafer damage by energetic photons from UV sources.

Various optical sources have been used in the commercial RTP systems [5.4.17] some of which are listed in the following:

- Tungsten-halogen lamps (those used in our work).
- Arc lamps (argon, mercury, xenon).
- Blackbody (graphite) heaters: System complexity because of the need for mechanical shutter and vacuum chamber is a major concern. Lack of feedback temperature control, possible contamination from graphite, and vacuum-only processing are additional major drawbacks.

Any possible radiation damage to the wafer by the energy source must be considered carefully. For instance, sources with high-energy UV photons, particle beams, and plasma may cause damage and device performance degradations. This problem does not exist in RTP systems with tungsten-halogen lamps or graphite heaters.

The optical radiation of the energy source reaches the wafer by traveling through a gas medium (unless vacuum processing is performed). Depending on the type of the process, the flowing ambient may consist of argon, nitrogen, oxygen, ammonia, forming gas, or some other gas. If the ambient gas has a strong absorption line or band falling within the source radiation spectrum, it can absorb a substantial amount of the optical energy in the form of gas heating and/or dissociation of the gas molecules and generation of active radicals. These undesirable phenomena may strongly affect the wafer temperature and the process kinetics. As an example, we have observed that a flowing ammonia ambient absorbs a significant amount of the optical radiation of some types of tungsten-halogen lamps during the nitridation processes. The undesirable effects appear in the form of slow temperature ramp-up and drifts, reduced maximum wafer temperature, and a heated gas ambient and quartz chamber causing slower ramp-down rates. It can also make the wafer temperature much more sensitive to the gas flow rate as a result of direct gas heating by the optical source (total heat loss would also be a function of the gas flow rate).

There are certain processes which require the presence of an ultrapure ambient of the process gas before ramping up the wafer temperature. A good example is the rapid thermal growth of thin gate dielectrics by oxidation and nitridation cycles. Reproducible nitridations demand an ultrapure ammonia ambient with less than 1 ppm of oxidant impurities. This requirement places a lower limit on the minimum purging time of the chamber with a flowing process gas (ammonia) after loading the wafer and before the temperature ramp-up. After loading the wafer, the chamber ambient consists of air with an atmospheric partial pressure (p_{air} = 1 atm). Then a constant flow (f in units of sccm) of the process gas enters the chamber inlet at one side while the outlet at the opposite side of the chamber directs the chamber gas with a similar flow rate into the exhaust. If the gas mixing inside the quartz chamber is assumed to occur rapidly, the time-dependent air content of the chamber in ppm is calculated from

$$p_{air} = \left(1\times10^{6}\right) e^{-(ft/V)} \ \ ppm, \tag{5.4.1}$$

where t and V are the purging time and the volume of the processing chamber, respectively. For example, the quartz chamber of the six-inch AET Addax 1000 rapid processor has a volume of

1685 cm^3. Therefore, with a maximum process gas flow rate of 5 *l/min* the worst case estimated air partial pressure remained in the chamber after 1, 2, 3, 4, and 5 minutes of purge will be 51100, 2650, 136, 7, and 0.36 ppm, respectively. This result implies that the RTP throughput for critical processes such as nitridation will be limited by the purging times and not the high-temperature cycle time itself (unless pumps and low-pressure chambers are employed). In some other processes such as annealing of tungsten, the processed wafer should not be unloaded unless the wafer temperature is lower than a critical value (less than about 200°C or otherwise air-induced oxidation will occur). One implication is that the wafer cooling rate may also influence the process throughput. However, proper design of the processing chamber (e.g. water-cooled stainless steel vacuum chambers) and addition of input/output vacuum loadlocks can significantly improve the throughput.

The optical source of an RTP system usually consists of an array of medium-power (about 2 kW per lamp maximum power) lamps (AET Addax, AG Associates, Varian) or a single high-power lamp (usually a line source $\geq$20 kW) with a focusing reflector (Peak Systems [5.4.18]). The heat loss at high temperatures in commercial RTP systems is mainly by radiation mechanism. Heat loss by convection becomes important for atmospheric high gas flow rates and during the ramp-down and when the wafer temperature approaches that of the chamber. Gas flows in the chamber during the inert or reactive ambient processing may cause nonuniformities caused by the convection gas cooling effects. At higher wafer temperatures, the radiative loss is several orders of magnitude higher than convection loss which is the reason for capability of fast wafer temperature cycling in an RTP system. To achieve fast ramp-down time it is essential to have a cold-wall system in which the surroundings are absorbing and reasonably cooled by air flow or water. Trapping of thermal energy in the quartz chamber caused by insufficient cooling raises its temperature and, therefore, severely restricts the throughput. Although the ramp-up of RTP can be very fast, but the temperature ramp-down may occur slowly when the chamber temperature is high. Some processes require postprocess purging in the process gas until the wafer is cold enough to be removed from its controlled ambient.

Following the above discussions, the important RTP equipment requirements for integrated circuit fabrication processes can be summarized as,

- uniform wafer temperature (less than ±5°C nonuniformity across the wafer)
- wafer holders with very low thermal mass
- reliable wafer temperature measurement and control (300°-1200°C)
- adequate chamber cooling for continuous high-temperature operation
- appropriate optical source for minimum gas absorption and negligible interference with the temperature measurement unit
- suitable wafer transport in and out of the process chamber for particulate control in a manufacturing environment
- negligible equipment-induced process contamination
- controlled ambients during heating, processing, and cooling and capabilities for reactive ambient processing
- programmable temperature and ambient gas cycles for multicycle *in-situ* processing.

As an example of equipment-induced process degradation, we have observed significant temperature uniformity degradations caused by the large quartz pins used as wafer holders in our AET Addax RTP equipment. These large wafer holders (three pins located circularly at 120° from each other) were causing localized cooling of the near-edge regions of the wafer in contact with the pins. Figure 5.5.1 plots the sheet resistance uniformity map of a 100 mm wafer rapidly annealed after implantation of singly ionized arsenic with a dose of 1.02×10^{15} cm^{-2} at 180 keV. The rapid thermal annealing conditions were: purge in high oxygen flow for 60 sec, ramp up to 1000°C in 15 seconds, perform rapid thermal annealing at 1000°C for 15 sec in oxygen ambient, and ramp down. This figure clearly shows the adverse effects of large thermal mass quartz pins on temperature uniformity through the sheet resistance map. The nonuniformity problem was because of significant localized heat loss by thermal conduction through the quartz pins and also shadowing of the lamp radiation by those large holders. These localized nonuniformities were eliminated by a significant reduction of the size and contact areas of the quartz pins used as wafer holders.

For successful implementation of the RTP technique into advanced fabrication processes several equipment and process parameters of practical importance must be taken into careful consideration. Besides automation, system throughput, and temperature measurement and control issues, process evaluations such as possible equipment-induced damage to wafer (formation of slip lines, temperature-induced stress, and/or radiation damage) and run-to-run reproducibility and uniformity are highly valuable. An equipment related example deals with the edge-induced wafer heat losses which prevent achieving perfect wafer temperature uniformity while a uniform optical flux distribution is incident on the wafer. With a uniform optical flux, the steady-state wafer temperature is gradually reduced from the center towards the edge. The steady-state spacial temperature gradients may be minimized by increasing the optical flux near the wafer edge. Although this solves the steady-state nonuniformity problem but it may actually worsen the transient temperature gradients during the ramp-up cycles. The dynamics of wafer edge heating during the ramp-up period of the process is different from those of the central region of the wafer. When the wafer temperature is quickly ramped up to a specified temperature by sensing and controlling the temperature of the wafer center in a feedback loop, the edge temperature experiences an overshoot at the end of the transient ramp-up time. Both the steady-state edge loss-induced temperature gnonuniformity and the wafer edge transient overshoots may result in large radial temperature gradients and, as a result, formation of slip lines. Wafer slip lines are actually defects extending from the edge of wafer caused by the shear displacements. It is desirable to minimize the RTP time at high process temperatures and improve the temperature uniformity to prevent slip formation. We have observed that slower temperature ramp-ups and ramp-downs and the use of large silicon trays can both help to eliminate the slip line formation by reducing the radial temperature gradients. To maintain a uniform steady-state wafer temperature, the edge region usually needs a larger optical power flux than the center. The use of a silicon ring surrounding the wafer is an effective approach to minimizing the extra edge-induced radiation loss. Another solution is to design the heat source (usually in single lamp systems) for a higher incident power near the wafer edge. This can be easily accomplished for blackbody heat source by varying the source configuration to obtain the desired spacial heat profile on the wafer. Moreover, the focusing reflectors of single lamp sources can be designed to increase the incident power at the wafer edge.

5.4.5 Temperature Measurement and Control

Because many of the processes performed by RTP are activated with temperature, accurate measurement and control of the wafer temperature is a necessity to reproduce a thermal cycle over many runs. Some previous works have addressed various sources of measurement errors in the measurement devices [5.4.19]. The wafer temperature can be measured either directly (contact) or indirectly (noncontact) by various techniques including,

- thermocouple directly attached to the wafer
- thermocouple attached to an underlying silicon tray
- thermocouple attached to an adjacent control silicon chip
- pyrometer [5.4.33]
- novel noncontact probing techniques by lasers, microwaves, ultrasonic waves, etc.

Wafer can be heated either in an open-loop (usually constant lamp power) or in a closed-loop (normally constant temperature or controlled ramps) mode. The former yields a slower temperature ramp-up and poor run-to-run process reproducibility. In contrast, the latter is very effective to produce fast temperature ramp-ups and reproducible heating cycles but could cause large transient radial temperature gradients (unless the equipment design is optimal). Thermocouples offer a fairly precise approach (within ±5°C uncertainty) to temperature measurement, however, they may not be extensively used in production systems because of strict manufacturing requirements. A thermocouple directly in contact with the device wafer can disturb the wafer temperature and cause some process nonuniformities unless it has a very low thermal mass. Large thermal mass thermocouples may result in temperature instabilities (ringing) while used in a feedback temperature control loop. We have observed that the chromel-alumel thermocouples less than 5 mils (0.127 mm) in diameter did not produce any ringing on fast temperature ramp-up cycles during RTP of 75 and 100 mm wafers.

The measurements taken by thermocouples connected to silicon monitor chips may not be an accurate duplicate of the actual wafer temperature and are susceptible to large errors (more than 50°C) depending on the differences between the silicon monitor chip and the actual process wafer. While the feedback controller achieves a perfect temperature and heating cycle for the silicon monitor chip, the temperature of the device wafer may have a slow ramp-up or experience a large overshoot. Thermocouples have a finite operational lifetime at high temperatures, particularly in reactive (i.e. oxidizing or nitridizing) ambients.

Based on our experiences with rapid thermal oxidation and nitridation processes, pyrometer is the most practical approach to RTP wafer temperature measurement. Although pyrometer appears as the most attractive RTP temperature measurement technique, there are some complicating factors which impose several limitations on its operation. Because the pyrometer measurement is also a function of surface emissivity which is not accurately known in advance of the measurement, the pyrometer should be frequently calibrated by using a thermocouple. Moreover, the pyrometer sensor may be easily disturbed by the optical radiation of the lamps and the heat trapped in the chamber. To obtain reproducible and accurate temperature measurement, the spectral interference or overlap between the pyrometer and the light source must be minimized. Further improvements should be expected if the wafer is viewed by the pyrometer at one side and is heated by the lamps

only at the opposite side so that pyrometer does not view the optical source directly. Pyrometer in a closed-loop temperature control can cause steady-state temperature drift with time if the wafer emissivity changes during the process (e.g. growth of oxide by rapid thermal oxidation or deposition of tungsten). The effect of chamber heating can be reduced by preheating the wafer (to, for example, 400°-750°C for a few seconds) before ramping to the process temperature; however, this preheating reduces the throughput because of slower ramp-down cycles caused by chamber heating. Chamber heating in RTP systems is avoided when water-cooled stainless steel chambers are used.

Wafer temperature uniformity and distribution can be determined by a variety of methods including,

- thin film thermocouple arrays
- scanning *in-situ* noncontact probes (laser or microwave beams)
- infrared imaging by CCD arrays (infrared camera)
- activation of implanted dopants
- resistivity reduction in silicides
- oxidation and nitridation
- reverse leakage current in implanted diodes

We have successfully employed the rapid thermal oxidation technique to characterize and evaluate the temperature uniformity of RTP systems with tungsten-halogen lamps. By designing appropriate temperature-time profiles and heating cycles it is possible to study the uniformities both for the steady-state and the transient ramp-up conditions.

5.4.6 Equipment Modeling

Advanced automated semiconductor manufacturing operations have the general objectives of enhanced wafer fabrication yield and throughput. Single-wafer rapid thermal/plasma processing meets the automation requirements quite well. In contrast to batch processing, single-wafer processing deals with wafer-to-wafer process uniformity as a run-to-run uniformity issue. As a result, the task of equipment modeling and identification and characterization of equipment and process correlations becomes more straightforward and time-invariant.

Our near future objectives for RTP equipment modeling will be accomplished on two RTP systems. One system already in use at CIS (AET Addax 1000) has been dedicated to the anneal and growth processes without any extensive *in-situ* process monitoring capabilities at present. The other system to be installed in the near future is a custom-made RTP thermal/plasma reactor for growth and deposition processes and with capabilities for addition of a variety of *in-situ* process monitoring tools. We intend to create the appropriate computer interface to both of these systems. The computer will have a database consisting of a variety of process recipes which can be loaded into the RTP machine for any desired process. The real-time process monitoring data (temperature distribution, gas flow pattern, pressure, plasma intensity, electron and ion energies and densities, optical intensity of the source lamps, thickness and resistivity data, ...) will be fed back and

collected by the computer. The same data can be employed to perform additional real-time process control by the computer, yield studies, and statistical process analysis.

Theoretical studies will be conducted to calculate the temperature profiles across silicon wafers for uniform and nonuniform planar (or meshed) optical sources during the transient and steady-state heating cycles. The effects of source radiation distribution profile and temperature ramp-up rate on maximum spacial temperature variation will be evaluated. Lateral temperature gradients resulting from the presence of microlithographic wafer patterns will also be estimated. Possible improvements in temperature measurement and uniformity by the use of pyrometers in conjunction with silicon/silicon carbide-coated graphite disks (emissivity near 1) and silicon rings will also be investigated. Effects of high gas flow rates on temperature uniformity caused by convection losses will be studied and modeled.

In the area of plasma deposition of tungsten and growth of dielectrics plasma wafer damage and process nonuiniformities will be correlated to the device yield and plasma conditions. The impact of plasma on wafer temperature and its uniformity will also be examined.

5.4.7 RTP Applications in Silicon IC Technologies

This section presents some important RTP applications in advanced MOS technologies. We have been conducting research in various RTP application areas including rapid dielectric growth on silicon, germanium, and gallium arsenide. Moreover, we have brought the RTP and microwave plasma technique together in a novel equipment design dedicated to rapid thermal/plasma deposition of tungsten and dielectric processing for *in-situ* MOS device fabrication and equipment modeling activities. Comprehensive studies of novel process applications and identification of important process parameters in each specific application are prerequisite to the successful development of application-specific RTP equipment models and yield studies. The RTP processes are performed in vacuum, inert, or reactive gas ambients. Vacuum processing may reduce the contanimation risk but can only apply to a limited number of applications. In contrast to vacuum or inert ambient processing, reactive gas ambients significantly enhance the RTP applications to virtually all growth and deposition processes. Some tasks may be performed in any of these ambients. An example is glass reflow [5.4.21] which can be done in nitrogen, steam, or vacuum.

Implanted Dopants

The junctions formed by ion implantation consist of electrically inactive dopant atoms and a large density of crystal defects. High-temperature annealing of implanted dopants by RTP does the following:

- Solid phase epitaxy (SPE): This is a fast regrowth of the amorphized region back into a single crystal phase. The SPE rates at T$\geq$900°C are quite large (more than 10 μm/sec) and, as a result, SPE in a 0.1 μm typical shallow junction is done in a fraction of a second.

- Dopant activation: Electrical activation is also a fast process and mostly occurs during SPE. It is increased at higher temperatures and , for example, anneals at T$\geq$1100°C can result in over an order of magnitude higher activation than at 800°C.

- Elimination of defects: This is a slower process than SPE and dopant activation.

- Dopant diffusion: Consists of the transient and normal diffusion components. The former happens very fast but the latter is a much slower process.

In ion implanted junctions, there is a trade-off between junction depth and defect density; higher annealing temperature results in smaller defect density (or lower junction leakage) but deeper junctions. RTP overcomes the limitations of this trade-off by reducing the dopant redistribution during high temperature anneals because of short processing times. At the initial rapid thermal annealing stage of boron and phosphorus (but not for arsenic) implanted junctions, there is a fast transient diffusion independent of the anneal temperature. This transient diffusion extends over ~400 Å for boron and sets a technological limit on scaled p-channel IGFETs in submicron CMOS with implanted junctions. For formation of ultrashallow junctions both the implantation-induced dopant channeling and initial diffusion transient must be minimized. These undesirable phenomena can be eliminated by an amorphization implant (such as Si implant) prior to the dopant implantation.

Rapid Thermal Oxidation and Nitridation

Rapid thermal processing of silicon in oxygen and ammonia ambients is an attractive technique for growth of thin dielectrics such as silicon nitride, silicon dioxide, nitrided oxides, oxidized nitrides, and application-specific dielectrics (such as oxides with a buried layer of nitride near the interface). Multicycle rapid thermal growth processes are suitable for dielectric engineering and *in-situ* formation of thin layered insulators with a variety of controllable oxygen and nitrogen compositional depth profiles by appropriate design of the temperature and ambient gas cycles.

Nitroxide films are prepared by rapid thermal nitridation (RTN) of SiO_2 films, usually grown in a furnace [5.4.11],[5.4.22]-[5.4.25]. Because the preparation of RTN nitroxide requires the growth of SiO_2 before RTN, the oxidized silicon wafers should be transferred from the furnace into the RTN chamber following oxidation. *In-situ* multiprocessing can enhance yield and reduce contamination and, as a result, it is advantageous to perform both the oxidation and nitridation processes via RTP which will greatly simplify the growth of RTN nitroxide. Conventionally, SiO_2 films have been grown in standard furnaces where oxidations are long ($t \geq 20$ min), and lower oxidant partial pressures are required to grow very thin films with good electrical characteristics.

To develop thin layers of silicon nitroxide by rapid thermal nitridation, silicon was first oxidized in a furnace to grow a thin (100 Å) layer of SiO_2; subsequently, the oxidized silicon wafers were subjected to ammonia at high temperatures (900° to 1200°C) to convert SiO_2 into layered silicon nitroxide. One motivation for promoting the rapid-thermal-oxidation (RTO) technique described in this section was the growth of initial SiO_2 by RTO instead of furnace oxidation prior to the RTN cycle. The necessary two-step processing in two different pieces of equipment would be eliminated which would greatly improve quality, yield, and process simplicity because RTN can then immediately follow RTO in the same RTP chamber to grow high-quality silicon-nitroxide films on silicon. Extensive new results obtained from the RTO process are reported here, including the initial regime of thermal oxidation of ⟨100⟩ silicon in dry O_2 and the SiO_2 films grown via this technique under a wide range of experimental growth conditions [5.4.13,5.4.14]. Investigation of thin SiO_2 growth kinetics in the very short oxidation regime with standard oxidation furnaces is not precise because the transient times involved in furnace processing are much longer than the very short oxidations to be studied; however, RTP has become a nearly ideal tool for these applications.

RTO Growth Kinetics Studies

In an earlier series of experiments [5.4.26], the RTO of ⟨100⟩ silicon was performed in a Heat-pulse 210-T RTP system. The 60 and 80 percent lamp-array power settings were estimated to correspond to constant RTO temperatures of 1040° and 1140°C, respectively; these experiments extended over 5 to 150 sec in dry O_2 (100% O_2) and in O_2-Ar mixtures (50% or 25% O_2). The growth kinetics were reported to be linear with time under all specified experimental conditions and the thickness data were well described by the Deal-Grove oxidation model [5.4.27]; however, the linear rate constants calculated from the data were significantly larger than those obtained by furnace oxidation at the same temperatures. At an estimated 1040° and 1140°C, for example, the reported growth rates were 1.27 and 2.34 Å/sec, respectively. The observed linear growth kinetics over a wide time range down to very short RTOs vary from the results obtained in this investigation [5.4.13,5.4.14] and in the experiment with thin-oxide growth in dry O_2. The thin-oxide data from those grown in furnaces down to the short times (a few minutes) achievable in furnaces indicated nonlinear growth kinetics in the initial regime of oxidation in a dry ambient [5.4.28] which is consistent with other observations of the anomalous behavior of dry oxidation in the thin-oxide regime whose physical mechanism is still a controversy. It has been proposed that, as silicon is oxidized in a dry O_2 ambient, a fixed positive charge forms near the oxide-silicon interface and reduces the density of holes at the Si-SiO_2 interface; the oxidation rate decreases as the charge layer forms because the density of broken Si-Si bonds at the interface is directly related to the hole density [5.4.29].

The observation of linear RTO growth kinetics [5.4.26] could have been the result of RTO temperature-estimation errors. The mapping of a constant lamp power to a constant temperature in an open-loop lamp-heated system (with no feedback temperature control) is not precise because a constant lamp power corresponds to a slowly rising wafer temperature as RTO time is increased. The thermal energy trapped in the chamber (including the quartz isolation chamber) and the finite loss rate of thermal energy from the chamber by radiation are the reasons why a constant lamp power corresponds to lower wafer temperatures in shorter RTOs and higher wafer temperatures in longer RTOs. As a result, the reported linear growth-kinetics data [5.4.26] are not isothermal curves vs RTO time.

In a second series of experiments [5.4.30], the RTO of epitaxial p-p^+ ⟨100⟩ silicon wafers was performed at 1150°C from 5 to 35 sec followed by N_2 annealing in a similar RTP system with feedback temperature control. Based on the measured thickness data, a constant growth rate of 3.3 Å/sec at 1150°C was reported for the limited RTO interval of 5 to 20 sec and a thickness uniformity of better than 2 percent on three-inch wafers.

The electrical characteristics of RTO SiO_2 films obtained in the first experiments [5.4.26] differed from those in the second [5.4.30]. The first series indicated moderately high fixed charge density at the interface and a faster generation rate of surface states by constant-current injection than in the furnace-grown SiO_2 and concluded that the electrical characteristics of RTO SiO_2 are inferior to the furnace-grown SiO_2 [5.4.26]. In the second series, however, the electrical breakdown characteristics of the RTO SiO_2 were observed to be equal to or better than the furnace-grown oxides [5.4.30]. These differences may be the result of processing variations and details of RTO and furnace SiO_2 growth, device fabrication, and system cleanliness, and firm conclusions regarding the RTO process cannot be reached without a large set of reliable RTO data.

The RTO experiments and kinetics data are described in the following sections, and the RTO

growth kinetics and existing thin-oxide models will be compared.

Experimental Procedure for Growth Kinetics Studies

The set-up used for the oxidation experiments was a lamp-heated Heatpulse 610 rapid thermal processor similar to the configuration used in the RTN experiments [5.4.11]. The temperature measurement was obtained by a pyrometer in a feedback loop for closed-loop temperature control. In each experiment and preceding the RTO runs, the calibration settings of the pyrometer were determined in dummy runs by comparing them to a thermocouple directly attached to a silicon wafer at all the temperatures to be applied in the following runs. In subsequent runs, only the pyrometer (with the correct calibration setting at each RTO temperature) was used for temperature measurement and control.

Phosphorus-doped ⟨100⟩ three-inch silicon wafers with resistivities ranging from 5 to 7.3 Ωcm were chemically cleaned (H_2SO_4:H_2O_2:H_2O, HCl:H_2O_2:H_2O, 50:1 H_2O:HF, NH_4OH:H_2O_2:H_2O) and rapidly oxidized in a pure dry O_2 ambient at 950° to 1200°C (steps of 50°C) for 5 to 300 sec.

During each RTO run, the cleaned Si wafer with a chemically grown native oxide of 15±0.6 Å (as measured by an ellipsometer) was loaded inside the quartz isolation chamber and the system was purged with a high flow of dry O_2 (~4 l/min for 2 min); then, setting the flow at 1 l/min, the temperature was quickly ramped up to the desired oxidation temperature. Following the RTO process for the specified oxidation time, the temperature was ramped down in pure dry O_2.

Because the ramp-ups and ramp-downs were short compared to the actual RTOs, their effect on the growth kinetics could be neglected; however, to study these effects, experiments were conducted at a given temperature and time with temperature-time profiles consisting of one and two consecutive heating cycles equal to the same total RTO time. There were two separate experimental sessions, and two different RTP systems were used to determine the system-to-system growth uniformity and the stability of temperature calibration. The pyrometer was calibrated in each session by a thermocouple and dummy runs at all the temperatures before the actual runs.

Additional RTO experiments were developed wherein the RTO process was followed with argon annealing at the oxidation temperature. The grown SiO_2 films were characterized with ellipsometry and by the electrical measurements of MOS capacitors fabricated with RTO oxides as gate insulators. Thickness was measured with 6328 Å ellipsometry, and the substrate and film refractive indices were set at 3.85-0.02i and 1.46, respectively. More than 20 measured points were distributed on each three-inch wafer to obtain the thickness uniformity across the wafer.

Growth Kinetics Results and Discussion

Figure 5.5.2 plots the RTO oxide thickness as a function of time at various temperatures (as measured by the pyrometer). Included are some experimental RTO data (solid and dash-dot curves hand fitted to the experimental data points) and SUPREM III [5.4.31] simulation results for dry oxidation from 950° to 1200°C (dashed curves). The height of the thin vertical line at each data point corresponds to twice the standard deviation of the thickness measured on the silicon wafer. The percentage ratio of standard deviation to average oxide thickness varied from system to system and ranged from less than 1 to 6 percent (mostly less than 4 percent). There did not appear to be a direct relationship between thickness uniformity and the RTO conditions because uniformity during both very short and long RTOs and at low and high temperatures were comparable and in the same range.

The experimental data obtained from two RTP systems in two experimental sessions were in agreement and reproducible at temperatures ranging from 950° to 1100°C (single solid curves in Fig. 5.5.2). The good growth-kinetics reproducibility in this range indicates that it is possible to calibrate the pyrometers for reproducible temperature vs time profiles in RTO experiments. The system-to-system RTO reproducibility (with pyrometer temperature measurement and control) appears to be acceptable at $T \leq 1100°C$. At 1150° and 1200°C, the dual curves in Fig. 5.5.2 correspond to the results obtained from two different rapid thermal processors although the pyrometer was calibrated similarly with the aid of a thermocouple in each. System-to-system thickness uniformity in the high-temperature range ($T \geq 1150°C$) was observed to be poor (~20 percent worst-case thickness variation) which may be attributed to the variations of temperature control in the systems containing pyrometers. The high optical intensity in the isolation chamber at elevated temperatures (corresponding to a large lamp optical output) and the hot-quartz isolation chamber can interfere with the pyrometer measurement of wafer temperature unless appropriate precautions are taken. This problem can be resolved with a pyrometer less sensitive to the background lamp-radiation spectrum.

The experimental data indicate the nonlinear growth of RTO oxide; its growth rate is the highest at the initial stage and is gradually reduced during longer RTOs. All curves in Fig. 5.5.2 start from 15 Å which is the thickness of the chemically grown native oxide. During the first 5 sec of the process, the initial growth rate R_0 can be calculated from the RTO data, and Fig. 5.5.3 plots this parameter on a logarithmic scale (averaged over the initial 5 sec) vs $1/kT$ corresponding to 950° to 1200°C.

The solid straight line is the least-square fit to the experimental data points. As can be seen, R_0 has a nearly Arrhenius dependence on temperature with an activation energy of 1.21 eV according to

$$R_0 \approx 1.81 \times 10^5 \, e^{-(1.21eV/kT)} \ \text{Å}/sec \, , \tag{5.4.2}$$

where $k = 8.62 \times 10^{-5}$ eV/K is Boltzmann's constant. The linear rate constant B/A in the Deal-Grove model for dry oxidation of ⟨100⟩ silicon has an activation energy of 2 eV [5.4.27]. Recent data, however, indicate two activation energies– – 1.76 eV at $T \leq 1000°C$ and 3.20 eV at $T \geq 1000°C$ [5.4.28] which are larger than 1.21 eV for R_0. This implies that, during the initial few seconds, RTO is less temperature dependent than the linear rate constant predicts which emphasizes the importance of an additional physical mechanism that affects the initial RTO regime. The ratio of R_0 to B/A is large and ranges from 37.4 to 9.5 at 950° to 1200°C, respectively.

The RTO oxides grown at 1200°C for 60 sec (one RTO cycle) and 20+40 sec (two consecutive RTO cycles) were 266.9 and 273.3 Å, respectively; a difference of 2.4 percent is the result of one additional temperature ramp-up and ramp-down and is indicative of their negligible effect on growth kinetics. No effect of argon annealing on the SiO_2 index of refraction (corresponding to oxide density) was observed.

By comparing the SUPREM III simulation results to the RTO experimental data, it can be observed that SUPREM III underestimates the initial growth rate by lowering it to between the experimental results and B/A values; it also underestimates SiO_2 thicknesses. Although a perfect agreement between these experimental data and the SUPREM III simulation results was not expected (because of the strong dependence of oxide growth rate on wafer cleaning procedures), the growth behavior of SiO_2 during the first few seconds of rapid oxidation appear to be different compared to the growth characteristics of longer oxidations. The existing more precise oxidation growth kinetics must be refined at least in the initial RTO regime where transient physical mechanisms

may play an important role.

Electrical and Reliability Properties

The kinetics of rapid thermal oxidation (RTO) of Si have been previously studied [5.4.14],[5.4.30],[5.4.32]-[5.4.34]. The preliminary electrical data for oxides grown by the RTO process have indicated 13 MV/cm breakdown fields with well-behaved conduction and C-V characteristics [5.4.30],[5.4.34]-[5.4.36]. This paper presents additional results on the ramped-voltage and time-dependent dielectric breakdown (TDDB) characteristics of MOS devices with rapidly grown unannealed gate oxides and the effects of preoxidation cleaning and native oxides on charge-to-breakdown, fixed oxide charges, and surface states.

MOS Device Fabrication

75-mm ⟨100⟩ n-type Si wafers with resistivities in the range of 0.1 to 0.9 Ωcm were cleaned and then oxidized at 1000°C to grow 0.74 μm of field SiO_2. After defining the gate areas by lithography and wet etching the wafers were divided into two groups. One group was cleaned using H_2SO_4:H_2O_2 (90°C), DI H_2O rinse, 50:1 H_2O:HF dip, DI H_2O rinse, HCl:H_2O_2:H_2O (70°C), DI H_2O rinse, NH_4OH:H_2O_2:H_2O (70°C), and DI H_2O rinse leaving a 15 Å-thick chemically grown native oxide. This native oxide is formed by the hydroxide/peroxide solution and not the final water rinse and its thickness measured by ellipsometry (index of refraction=1.46) is in close agreement with the cross-sectional TEM results. The other group was cleaned similarly but with a final 50:1 H_2O:HF dip and DI H_2O rinse after the hydroxide/peroxide cleaning to remove the native oxide; however, a very thin (~5 Å measured by ellipsometer) layer of native oxide is formed on the surface of freshly etched and DI water-rinsed Si wafers unless an *in-situ* surface cleaning is employed. The ellipsometry data indicated the differences between these two cleaning procedures because of the different native oxide thicknesses, although both cleanings had a final DI water rinse.

The RTOs were performed in a Heatpulse 610 system at 950°C for 300 sec, at 1000°C for 90 sec, at 1050°C for 30 sec, at 1100°C for 15 sec, and at 1150°C for 5 sec. The oxidation times were chosen so that the thicknesses of oxides grown at various temperatures were the same and about 80 Å. The wafer temperature was measured with a pyrometer and controlled in a feedback loop. The pyrometer was calibrated only at one temperature using a thermocouple. No attempt was made to calibrate the pyrometer at various oxidation temperatures and identical calibration settings were used during all the experiments. To preserve the intrinsic properties of rapidly grown oxides, no postoxidation inert ambient or final forming gas anneals were conducted (unless specified otherwise), although this would imply the presence of a higher density of fixed charges (Q_f) and surface states (D_{it}). The oxidized wafers were immediately transferred to a LPCVD furnace for deposition of 3500 Å polysilicon which was subsequently doped with phosphorus in $POCl_3$ at 900°C. The lower doping temperature of 900°C (compared to the more conventional 950°C) was chosen to ensure negligible phosphorus diffusion into and through the ultrathin oxides and, as a result, preserve their intrinsic properties. The polysilicon gates were defined by microlithography and plasma etching. After backside etching of polysilicon and oxide, 600 Å tungsten was deposited on exposed silicon regions (polysilicon gates and pads and backside of the wafers) by a selective LPCVD process in a flowing WF_6/H_2 ambient at 300°C. The comparisons between the device characteristics measured before and after tungsten deposition indicated negligible *in-situ* oxide

annealing by hydrogen at 300°C during the tungsten deposition process, however, the result could be different if a thinner layer of gate polysilicon was used.

Electrical Characterization

Oxide thicknesses were obtained by ellipsometery and electrical measurements of accumulation capacitances. When the n-type substrate is in accumulation, the n^+ polysilicon gate is depleted. Because the heavily doped polysilicon depletion layer width is much smaller than the oxide thicknesses, and the dielectric constant of Si is larger than that of SiO_2, the accumulation capacitance is equal to the SiO_2 capacitance. The pyrometer accuracy during the RTO experiments was not cross-examined by thermocouples at all temperatures, therefore, some temperature measurement error may be expected. As shown in Fig. 5.5.4, the oxides grown at 950°C were nearly 100 Å and those grown at 1000° to 1150°C were scattered around 80 Å with the exception of one data point (probably a result of a temperature calibration problem). The oxides grown on two groups of wafers with different cleanings had comparable thicknesses (less than 8 Å difference). Because of scattering of the measured thicknesses at various growth temperatures, no attempt was made to examine the dependence of SiO_2 thickness on surface cleaning, although differences of 8 Å are significant in the ultrathin regime.

An example of the C-V characteristics of MOS devices with 79 Å of as-grown RTO oxide is shown in Fig. 5.5.5(a). The high- and low-frequency C-V characteristics of MOS devices with unannealed oxides were somewhat distorted because of surface states which could be annealed out by postoxidation and forming gas anneals. For example, in MOS devices with 77 Å SiO_2 grown at 1150°C with a preoxidation HF dip, an Ar anneal at 1038°C for 45 sec (after polysilicon deposition and doping) reduced the midgap D_{it} from an average of 6.88×10^{11} to 1.49×10^{11} $cm^{-2}eV^{-1}$. Using the C-V data, D_{it} vs surface potential was calculated for a variety of devices and a typical result is plotted in Fig. 5.5.5(b) indicating a minimum D_{it} of 3×10^{11} $cm^{-2}eV^{-1}$ (midgap D_{it}=7.13×10^{11} $cm^{-2}eV^{-1}$). The C-V and D_{it} characteristics of all MOS devices with oxides grown at various temperatures were similar in general behavior; however, the key parameters such as flatband voltage (V_{fb}), and minimum and midgap D_{it}'s were dependent on the preoxidation cleaning and growth conditions.

Figure 5.5.6(a) plots the dependence of V_{fb} on RTO temperature and preoxidation cleaning. In the devices with HF dip, V_{fb} became more negative with higher temperature; this is attributed to a larger Q_f in oxides grown at higher temperatures because of a faster growth rate. On the other hand, the MOS devices without the preoxidation HF dip did not exhibit a similar trend and their V_{fb}'s were nearly independent of growth temperature. The more positive V_{fb} in the latter is because of a smaller Q_f when no HF dip is used. The effect of preoxidation cleaning on V_{fb} becomes less significant at lower temperatures where both cleanings result in near-zero V_{fb}'s. Since for a given Q_f, V_{fb} is also a function of gate oxide thickness (or gate capacitance C_{ox} according to $V_{fb} = \Phi_{ms}+Q_f/C_{ox}$, Φ_{ms} ≈-0.15 to -0.2 eV), the comparison of V_{fb}'s will be more meaningful if they are normalized to the gate oxide thickness. This is because of the slight variations in thicknesses of oxides grown at different temperatures. Any change in $V_{fb}\times C_{ox}$ should correspond to a direct change in Q_f (i.e. $\Delta Q_f = C_{ox}\Delta V_{fb}$) independent of the value of Φ_{ms}. A plot of $V_{fb}\times C_{ox}$ vs RTO temperature and shown in Fig. 5.5.6(b) resulted in conclusions which were consistent with those of Fig. 5.5.6(a). Because of the presence of chemical oxide on Si prior to RTO, Q_f in MOS devices without the preoxidation HF dip is found to be nearly independent of the growth temperature.

The minimum D_{it}'s in MOS structures with the two different preoxidation cleanings are shown in Fig. 5.5.6(c) as a function of RTO temperature. In the wafers with the final HF dip, $D_{it,min}$ decreases as the RTO temperature is raised but it appears to be temperature-independent when no final HF dip is performed. The effect of preoxidation cleaning on $D_{it,min}$ is negligible in the high temperature regime where D_{it}'s are minimum and converge together for both the cleanings.

As illustrated in Fig. 5.5.7, the I-V characteristics indicated a Fowler-Nordheim (F-N) tunneling mechanism over more than seven decades of current. The F-N parameters (preexponential and exponent) depend on the triangular barrier height (Φ_B) and the effective electronic mass in the tunneling distance of SiO_2 (m^*) and can be precisely determined. Since there are two equations and two unknowns (Φ_B and m^*), they both are calculated independently by fitting the data to the F-N equation. The calculated Φ_B and m^* (for devices with a final HF dip and 79 Å SiO_2 grown at 1050°C) were found to be 3.29 eV and $0.37m_0$, respectively (m_0=electron rest mass).

The ramped-voltage breakdown characteristics of MOS devices with oxides grown at various temperatures were investigated. Curves denoted as 'a' in Fig. 5.5.8 are measurements on different test chips of a wafer containing MOS devices with 79 Å SiO_2 and HF dip. The average destructive breakdown field (E_{bd}) where current abruptly increases is 15 MV/cm ($V_{fb} \approx 0$) and the characteristics exhibit good uniformity. At a voltage ramp rate of 1 V/sec no destructive breakdown was observed unless the current density exceeded 25 A/cm^2. Comparison of curves 'b' and 'c' reveals the effect of cleaning on E_{bd} and F-N conduction. These are for 100 μm^2 MOS capacitors with 105 and 98 Å oxides grown at 950°C on wafers with and without HF dip. In both cases the average E_{bd} is 15 MV/cm and E_{bd} and its integrity appear to be independent of the preoxidation cleaning; however, the F-N conduction distribution is tighter when HF dip is done, possibly because of the nonuniformities associated with the chemically grown native oxides.

The TDDB and trapping phenomena were investigated by the constant-current stress technique in which the gate voltage is monitored while a current is injected into the gate (substrate in accumulation). The change in constant-current voltage during the charge injection is related to charge trapping in the SiO_2 and the total charge density injected through the oxide up to the onset of destructive breakdown (Q_{bd} or charge-to-breakdown) is a measure of oxide quality. Figure 5.5.9(a) plots the voltage under constant-current stress TDDB at a stress current density of 1 A/cm^2 in 100 μm^2 MOS devices with 79 Å oxide grown at 1050°C on wafers with final preoxidation HF dip. The increase in constant-current voltage is an indication of net electron trapping which was observed to be much less in rapidly grown oxides compared to the oxides grown in a furnace. For example, in devices with 79 Å SiO_2 grown at 1050°C the rise in constant-current voltage up to the onset of destructive breakdown is less than 0.7 V and the average Q_{bd} is over 50 C/cm^2 at 1 A/cm^2, which is larger than 20 C/cm^2, the typical value measured for a furnace-grown 100 Å SiO_2 under similar stress conditions [5.4.37]. Figure 5.5.9(b) plots Q_{bd} (at 1 A/cm^2) vs RTO temperature for two different cleanings in 100 μm^2 MOS devices. The data indicate a higher Q_{bd} when the chemical oxide is removed by HF dip. As shown by the data in Fig. 5.5.10(a), Q_{bd} is significantly increased (over 100 C/cm^2) at lower currents because of a lower oxide electric field. In 10 μm×10 μm unannealed MOS devices with 79 Å SiO_2 grown at 1050°C with HF dip, the measured Q_{bd} data and their dependence on current density J could be least-square fitted to $Q_{bd} = Q_c ln(J_0/J)$ with the parameters Q_c = 18.94 C/cm^2 and J_0 = 23.92 A/cm^2. According to Fig. 5.5.10(b), Q_{bd} at a constant current density decreases with an increase in the gate area, as a result of the physical defects present in these oxides and the dominance of exterinsic defect related breakdown and wearout phenomena in the large area devices.

5.4.8 Summary

In conclusion, rapid thermal processing appears to be a viable multipurpose technique which can be effectively employed in many stages of advanced MOS processes. Our future research activities in this area will focus on equipment and process modeling issues, enhanced manufacturing capabilities via real-time *in-situ* process monitoring, and performance characterization and long-term reliability of RTP-based CMOS devices.

Rapid thermal oxidation of (100) silicon in a dry-oxygen ambient has been performed in a lamp-heated RTP system. Extensive new experimental results have been obtained from the initial regime of thermal oxidation of silicon by the RTO technique. These results reveal the nonlinear growth kinetics of rapid thermal oxidation in the short regime in contrast to recently reported data in the literature that suggest linear growth kinetics. The kinetics data show an increasing growth rate as RTO is reduced; the highest oxidation rate is for the shortest oxidation time. A very fast initial growth rate (followed by a reduction with RTO time or SiO_2 thickness) has an activation energy of 1.21 eV. Simple extrapolation of the existing long-time thin-oxide models to the short-time rapid-thermal-oxidation regime does not yield a precise prediction of the kinetics because different transient physical processes (other than simple one-species oxidant diffusion and interface reaction) may play an important role. The RTO process appears to be a very attractive and viable technique for growing good-quality thin SiO_2 for device applications either directly or after subsequent rapid thermal nitridation to form layers of silicon nitroxide.

According to our recent results, the tungsten/n^+ polysilicon gate MOS devices with rapidly grown unannealed SiO_2 have larger charge-to-breakdown and lower electron trapping compared to furnace oxides. The preoxidation cleaning affects the densities of fixed charges and surface states. As compared to the MOS devices with preoxidation HF dip, the presence of a chemical native oxide prior to RTO appears to reduce the fixed charge and surface-state densities, their dependencies on the RTO temperature, and the breakdown charge densities.

5.4.9 References

[5.4.1] P. Burggraaf, "Rapid wafer heating: Status 1983," *Semiconductor International*, pp. 68-74, Dec. 1983.

[5.4.2] J. Narayan, O. Holland, R. Eby, J. Wortman, V. Ozguz, and G. Rozgonyi, "Rapid thermal annealing of arsenic and boron implanted silicon," *Appl. Phys. Lett.*, vol. 43, no. 10, pp. 957-959, 1983.

[5.4.3] T. Yachi, "Formation of a $TiSi_2/n^+$ poly-Si layer by rapid lamp heating and its application to MOS devices," *IEEE Electron Dev. Lett.*, vol. EDL-5, no. 7, pp. 217-220, 1984.

[5.4.4] T. Hara, H. Suzuki, and M. Furukawa, "Reflow of PSG layers by halogen lamp short duration heating technique," *Jpn. J. Appl. Phys.*, vol. 23, no. 7, pp. L452-L454, 1984.

[5.4.5] T. Faith and C. Wu, "Elimination of hillocks on Al-Si metallization by fast-heat-pulse alloying," *Appl. Phys. Lett.*, vol. 45, no. 4, pp. 470-472, 1984.

[5.4.6] M. Reed, B. Fishbein, and J. Plummer, "Rapid thermal annealing of interface states in aluminum gate metal-oxide-silicon capacitors," *Appl. Phys. Lett.*, vol. 46, no. 4, pp. 400-402, 1985.

[5.4.7] Z. Weinberg, D. Young, J. Calise, S. Cohen, J. DeLuca, and V. Deline, "Reduction of electron and hole trapping in SiO_2 by rapid thermal annealing," *Appl. Phys. Lett.*, vol. 45, no. 11, pp. 1204-1206, 1984.

[5.4.8] M. M. Moslehi and K. C. Saraswat, "Thermal nitridation of Si and SiO_2 for VLSI," *IEEE Trans. Electron Devices*, vol. ED-32, no. 2, pp. 106-123, 1985.

[5.4.9] M. M. Moslehi, C. Y. Fu, and K. C. Saraswat, "Thermal and microwave nitrogen plasma nitridation techniques for ultrathin gate insulators of MOS VLSI," *1985 Symp. VLSI Technol. Dig. Tech. Papers*, pp. 14-15, (Japan) 1985.

[5.4.10] M. M. Moslehi, K. C. Saraswat, and S. Shatas, "Rapid thermal nitridation of Si and SiO_2 in ammonia," *1985 Electronic Materials Conf. Tech. Abstracts*, pp. 43-44, (Colorado) 1985.

[5.4.11] M. M. Moslehi, K. C. Saraswat, and S. C. Shatas, "Rapid thermal nitridation of SiO_2 for nitroxide thin dielectrics," *Appl. Phys. Letters.*, vol. 47, no. 10, pp. 1113-1115, 1985.

[5.4.12] M.M. Moslehi and K.C. Saraswat, "Rapid thermal nitridation of SiO_2 for nitroxide thin dielectrics," *Silicon Interface Specialties Conference (IEEE-SISC)*, (Florida) 1985.

[5.4.13] M. M. Moslehi, S. C. Shatas, and K. C. Saraswat, "Rapid thermal oxidation and nitridation silicon," *The Fifth International Symp. on Si Materials Sci. and Technol. (Electrochem. Soc.)*, (Boston) 1986.

[5.4.14] M. M. Moslehi, S. C. Shatas, and K. C. Saraswat, "Thin SiO_2 insulators grown by rapid thermal oxidation of silicon," Appl. Phys. Lett., vol. 47, no. 12, pp. 1353-1355, 1985.

[5.4.15] M. M. Moslehi, K. C. Saraswat, and S. C. Shatas, "Rapid thermal growth of thin insulators on Si," (Invited) *Proc. SPIE on Advanced Processing and Characterization of Semiconductors III*, vol. 623, pp. 92-114, 1986.

[5.4.16] M. M. Moslehi, S. C. Shatas, K. C. Saraswat, and J. D. Meindl, "Interfacial and breakdown characteristics of MOS devices with rapidly grown ultrathin SiO_2 gate insulators," Submitted to *Appl. Phys. Letters.*

[5.4.17] S. Wilson, R. Gregory, and W. Paulson, "An overview and comparison of rapid thermal processing equipment: A users viewpoint," *Mat. Res. Soc. Symp. Proc. on RTP*, vol. 52, pp. 181-190, 1986.

[5.4.18] T. Stultz, "Impact of rapid thermal processing on device technology," *Semicon/West Tech. Proc.*, pp. 87-94, 1986.

[5.4.19] R. Sheets, "Temperature measurement and control in a rapid thermal processor," *Mat. Res. Soc. Symp. Proc. on RTP*, vol. 52, pp. 191-197, 1986.

[5.4.20] J. Gelpey, P. Stump, and J. Smith, "Process control for a rapid optical annealing system," *Mat. Res. Soc. Symp. Proc. on RTP*, vol. 52, pp. 199-207, 1986.

[5.4.21] N. Shah, J. McVittie, N. Sharif, J. Nulman, and A. Gat,"Characterization of PSG films reflowed in steam using rapid thermal processing," *Mat. Res. Soc. Symp. Proc. on RTP*, vol. 52, pp. 233-240, 1986.

[5.4.22] C. Chang, A. Kamgar, and D. Kahng, "High temperature rapid thermal nitridation of silicon dioxide for future VLSI applications," *IEEE Electron Dev. Lett.*, vol. EDL-6, no. 9, pp. 476-478, 1985.

[5.4.23] J. Nulman, J. Krusius, and L. Rathbun, "Electrical and structural characteristics of thin nitrided gate oxides prepared by rapid thermal nitridation," *IEDM Tech. Dig.*, pp. 169-172, 1984.

[5.4.24] J. Nulman and J. Krusius, "Rapid thermal nitridation of thin thermal silicon dioxide films," *Appl. Phys. Lett.*, vol. 47, no. 2, pp. 148-150, 1985.

[5.4.25] W. Bailey, R. Sundaresan, and M. Matloubian, "Investigation of nitrided silicon dioxides for radiation tolerant MOS devices," *Silicon Interface Specialties Conference (IEEE-SISC)*, (Florida) 1985.

[5.4.26] G. Grant, G. Brown, J. Shu, E. Lee, and J. Reynolds, "Rapid growth of thin film silicon oxides," *The Electrochem. Soc. Meeting*, (New Orleans) 1984.

[5.4.27] B. Deal and A. Grove, "General relationship for the thermal oxidation of silicon," *J. Appl. Phys.*, vol. 36, p. 3770, 1965.

[5.4.28] H. Massoud, J. Plummer, and E. Irene, "Thermal oxidation of silicon in dry oxygen," *J. Electrochem. Soc.*, vol. 132, no. 7, p. 1745, 1985.

[5.4.29] S. Schafer and S. Lyon, "New model of the rapid initial oxidation of silicon," *Appl. Phys. Lett.*, vol. 47, no. 2, pp. 154-156, 1985.

[5.4.30] J. Nulman, J. Krusius, and A. Gat, "Rapid thermal processing of thin gate dielectrics. Oxidation of silicon," *Electron Dev. Lett.*, vol. EDL-6, no. 5, pp. 205-207, 1985.

[5.4.31] C. Ho and S. Hansen, "SUPREM III—A program for integrated circuit process modeling and simulation," Technical Report No. SEL 83-001, Stanford University, July 1983.

[5.4.32] A. Hodge, C. Pickering, A. Pidduck, and R. Hardeman, "Silicon oxidation by rapid thermal processing," *Mat. Res. Soc. Symp. Proc. on RTP*, vol. 52, pp. 313-319, 1985.

[5.4.33] J. Gelpey, P. Stump, and R. Capodilupo, "Oxide growth using the water-wall arc lamp," *Mat. Res. Soc. Symp. Proc. on RTP*, vol. 52, pp. 321-325, 1985.

[5.4.34] Y. Sato and K. Kiuchi, "Oxidation of silicon using lamp light radiation," *J. Electrochem. Soc.*, vol. 133, no. 3, pp. 652-654, 1986.

[5.4.35] C. Gronet, J. Sturm, K. Williams, and J. Gibbons, "Limited reaction processing of silicon: Oxidation and epitaxy," *Mat. Res. Soc. Symp. Proc. on RTP*, vol. 52, pp. 306-312, 1986.

[5.4.36] Z. Weinberg, T. Nguyen, S. Cohen, and R. Kalish, "SiO_2 growth and annealing by lamp heating," *Mat. Res. Soc. Symp. Proc. on RTP*, vol. 52, pp. 327-332, 1985.

[5.4.37] M. M. Moslehi and K. C. Saraswat, "Studies of trapping and conduction in ultrathin SiO_2 gate insulators," *IEDM Tech. Dig.*, pp. 157-160, 1984.

5.5 Figures

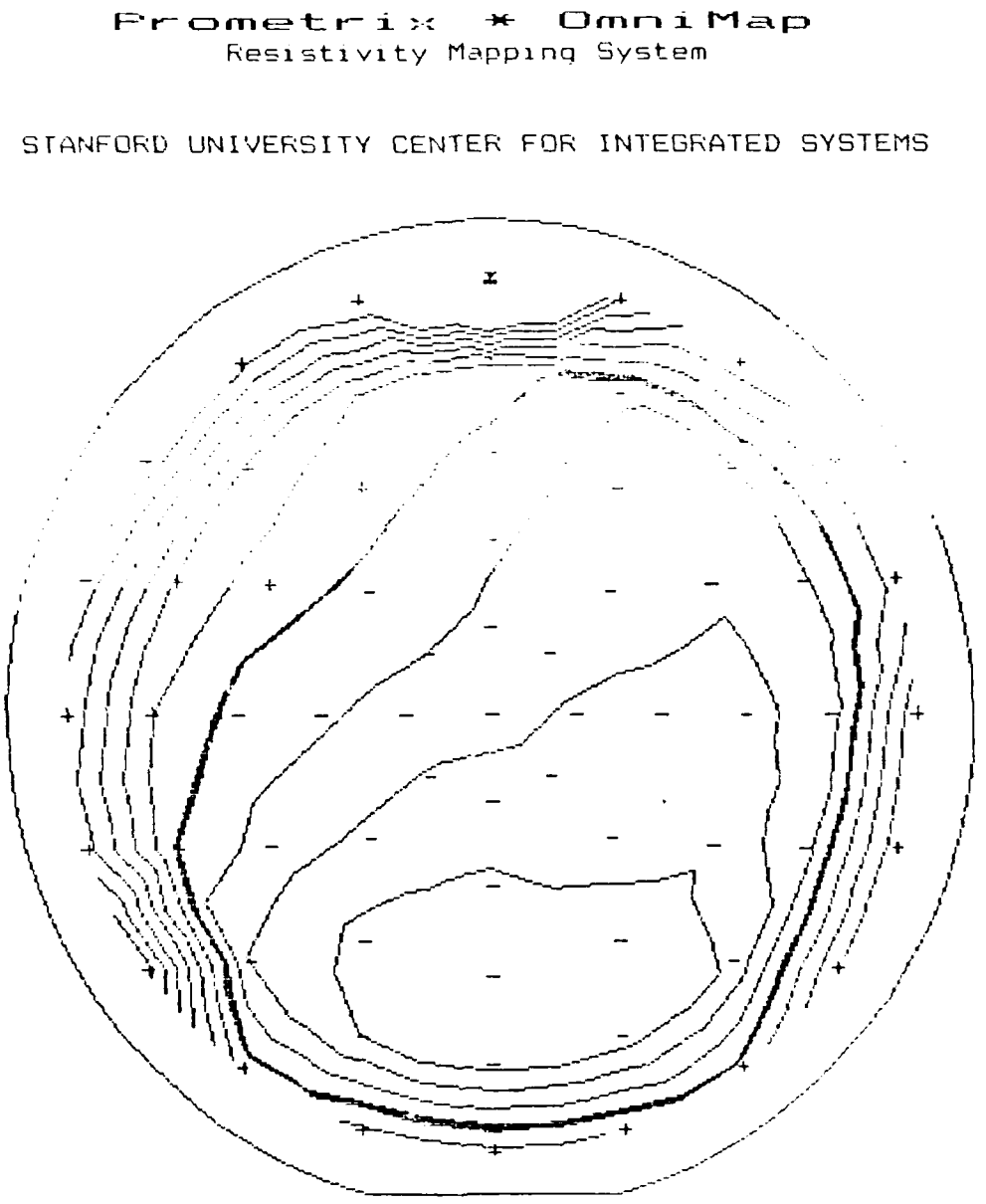
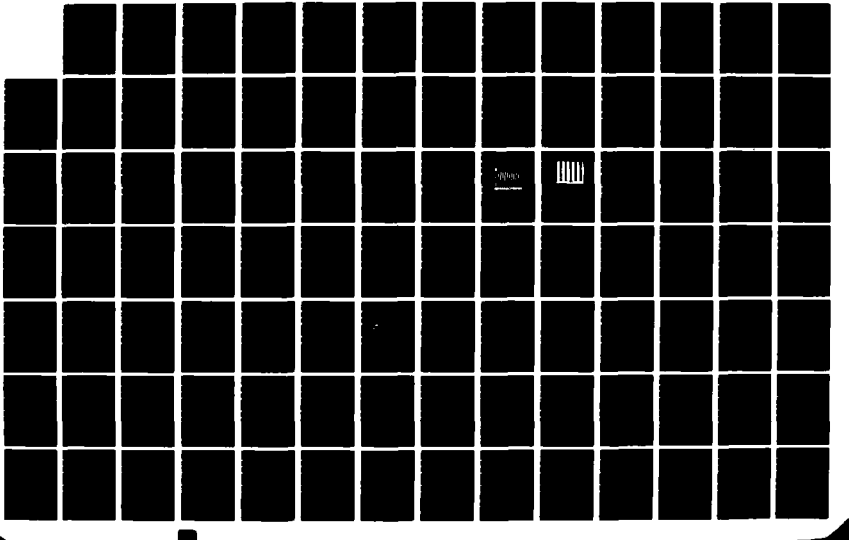
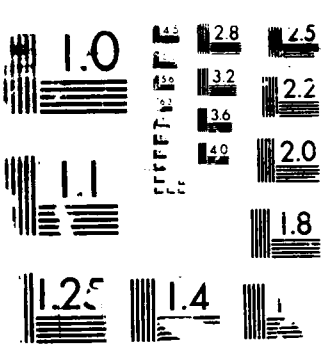


Figure 5.5.1: SHEET RESISTANCE UNIFORMITY MAP
SENIC-IMPLANTED 100 mm WAFER (average
tion=2.84%, contour interval=1.00%, test
at 180 keV).

AD-A189 451 COMPUTER AIDED FAST TURNAROUND LABORATORY FOR RESEARCH IN VLSI (VERY LARG (U) STANFORD UNIV CA CENTER FOR INTEGRATED SYSTEMS J D MEINDL ET AL 31 MAY 87 2/3

UNCLASSIFIED MDA903-84-K-0062 F/G 9/1 NL





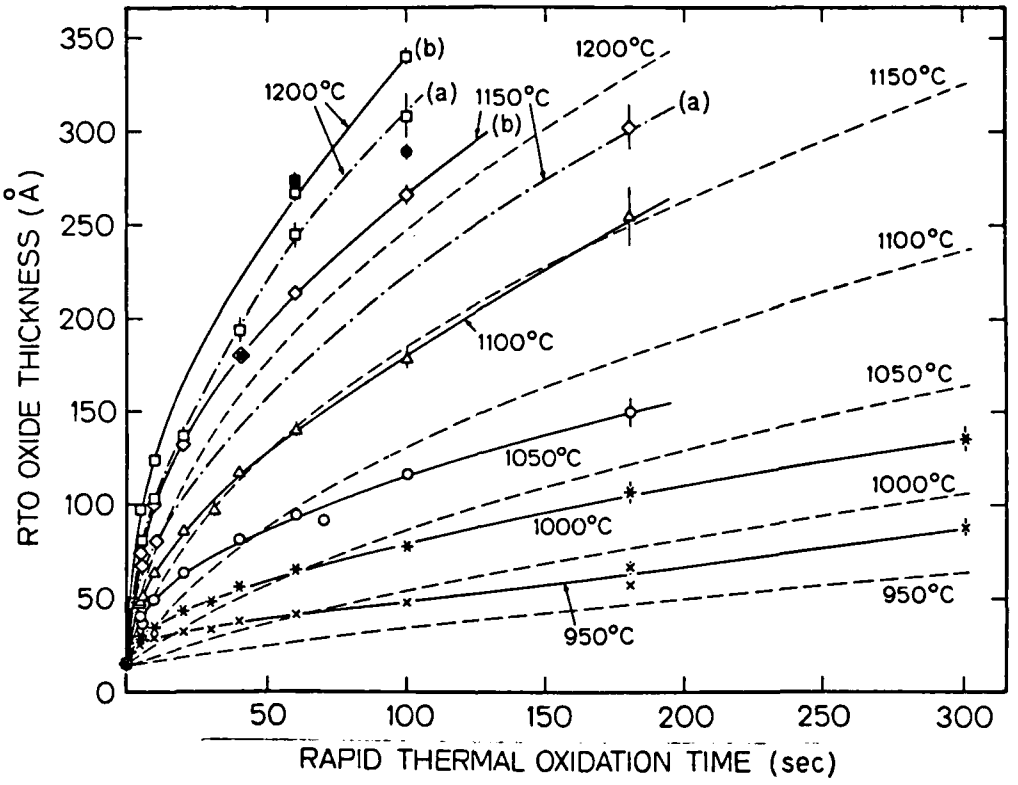


Figure 5.5.2: RTO OXIDE THICKNESS VS TIME AT VARIOUS TEMPERATURES. Solid and dash-dot curves were fitted to the data points and the dashed curves are SUPREM III simulation results. The symbols represent the data obtained under the following RTO conditions.

× :950°C
⋆ :1000°C
o :1050°C
△:1100°C
◇: 1150°C
□: 1200°C
• :1150°C, 40 *and* 100 *sec RTOs followed by* 30 *sec argon anneal*
□: 1200°C *two − cycle* (20 + 40 *sec*) *RTO*

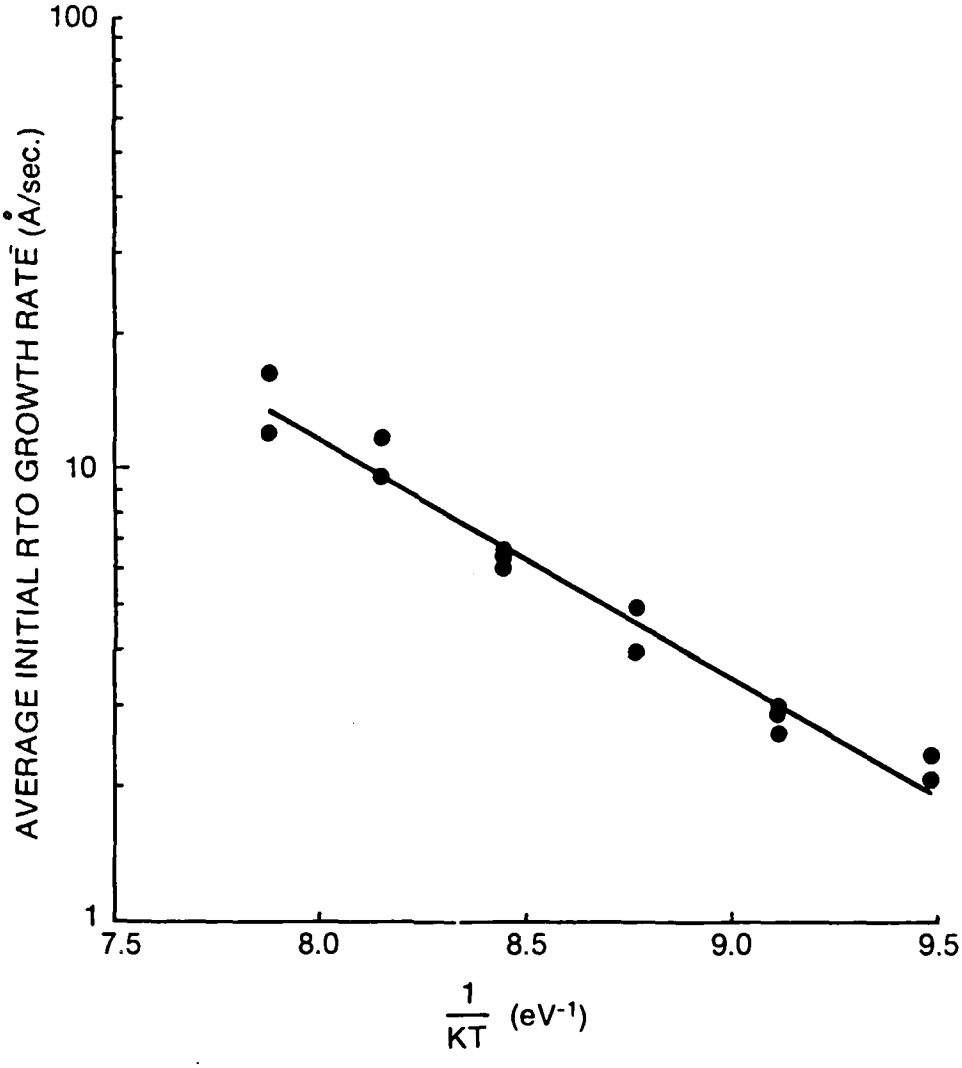


Figure 5.5.3: INITIAL RTO GROWTH RATE (averaged over the first 5 sec) VS $1/kT$ IN THE RTO RANGE OF 950° TO 1200°C. Activation energy=1.21 eV.

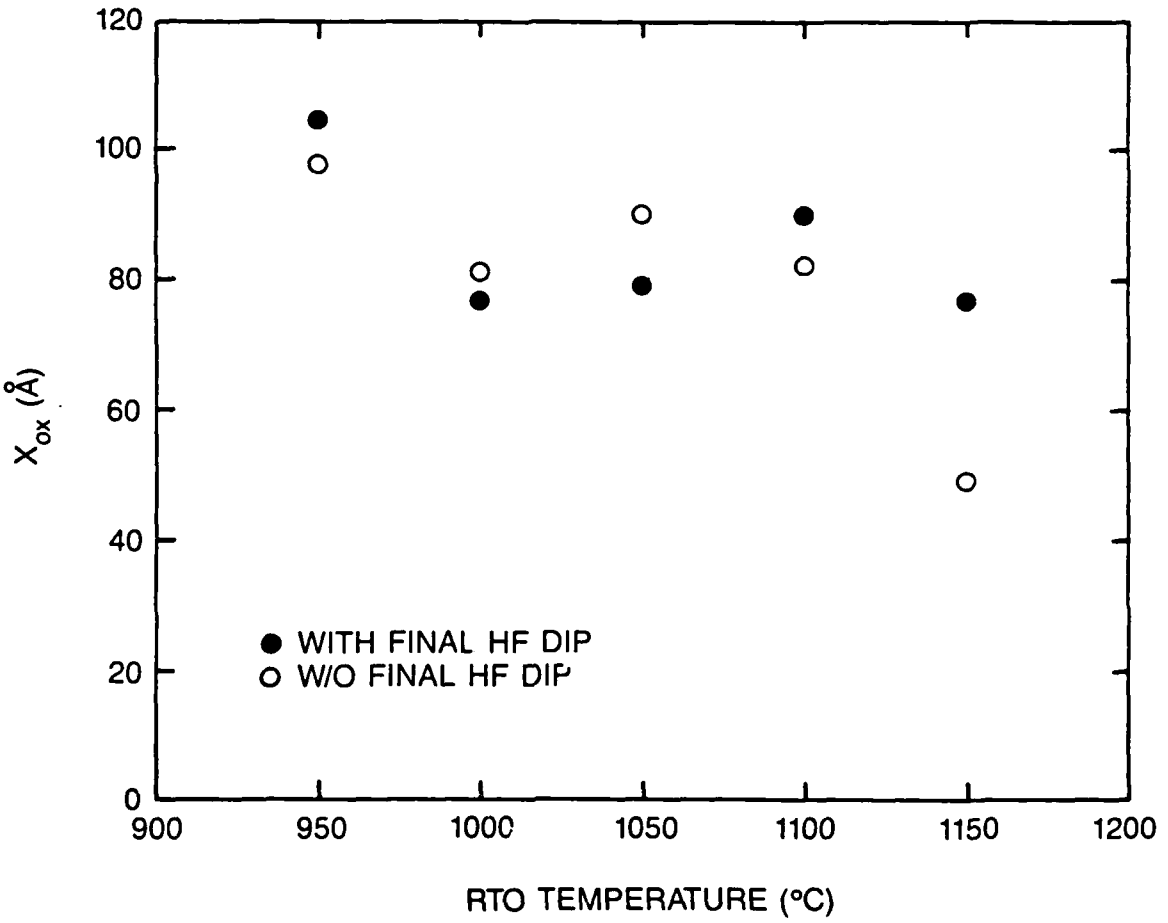


Figure 5.5.4: AVERAGE OXIDE THICKNESSES GROWN AT VARIOUS RTO TEMPERATURES determined from electrical accumulation capacitance measurements.

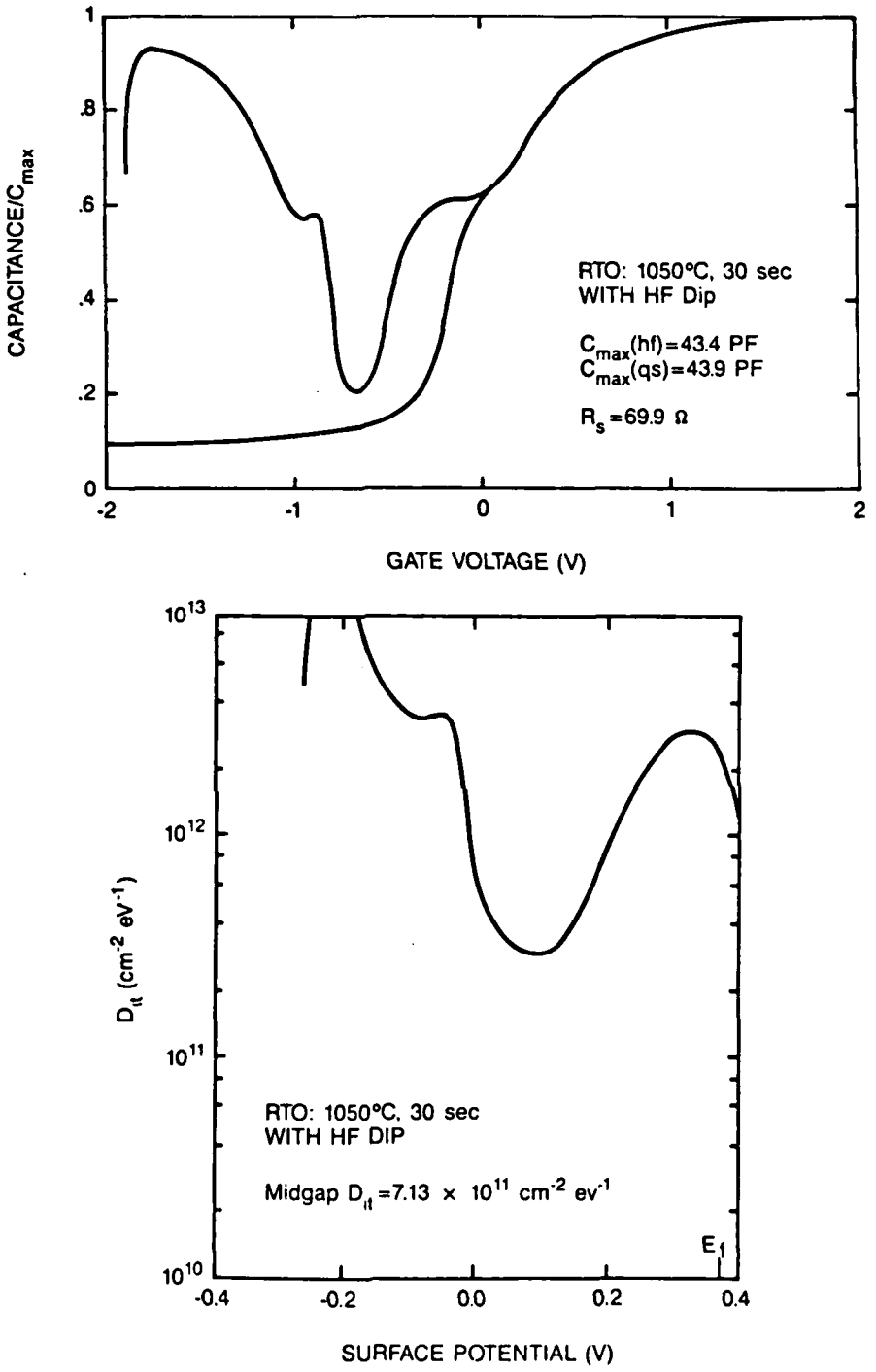


Figure 5.5.5: (a) HIGH- AND LOW-FREQUENCY C-V CHARACTERISTICS AND (b) SURFACE-STATE DENSITY VS SURFACE POTENTIAL in MOS devices with 79 Å SiO_2 grown at 1050°C on n-type Si with final preoxidation HF dip (gate area=100 $\mu m \times 100$ μm).

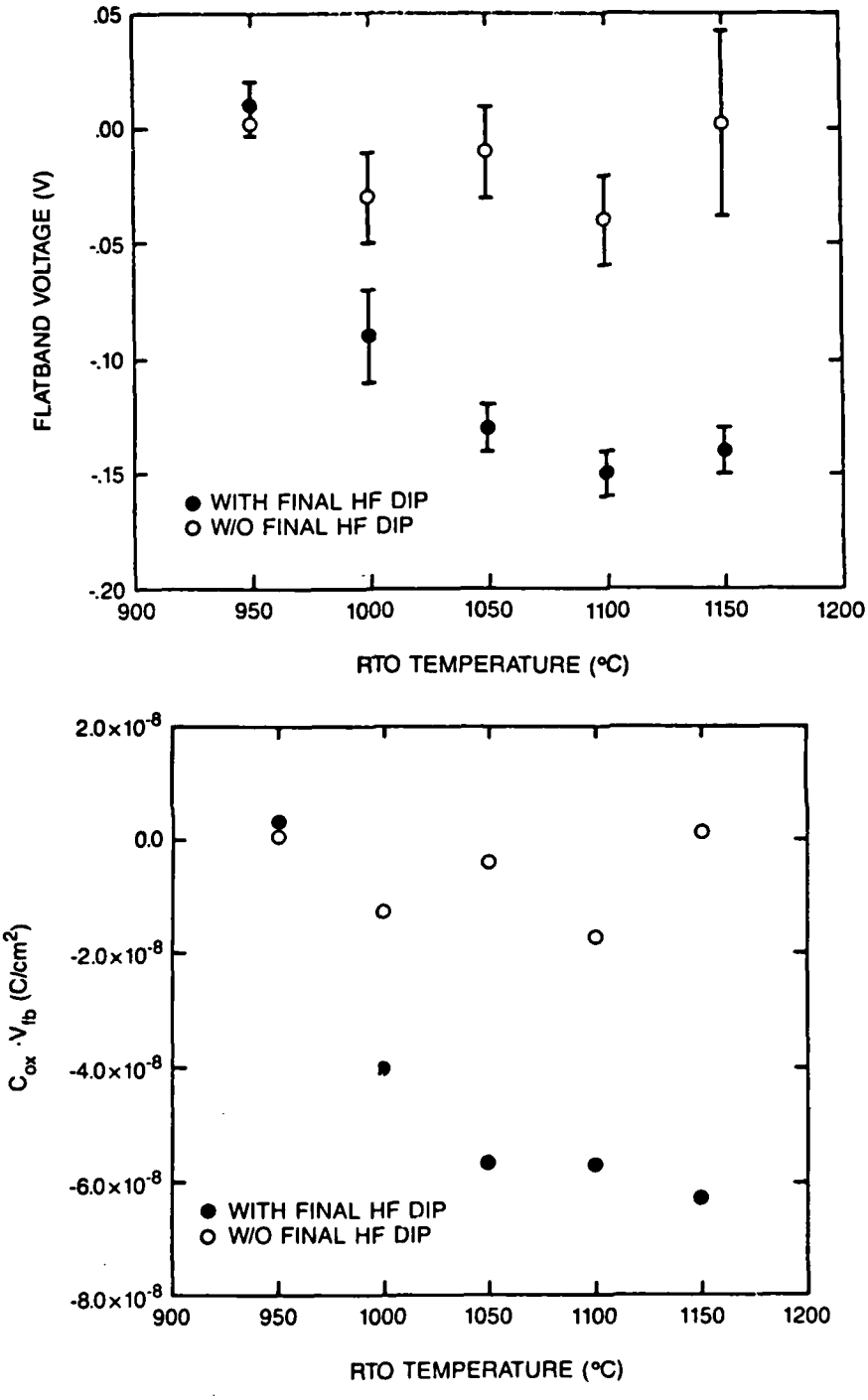
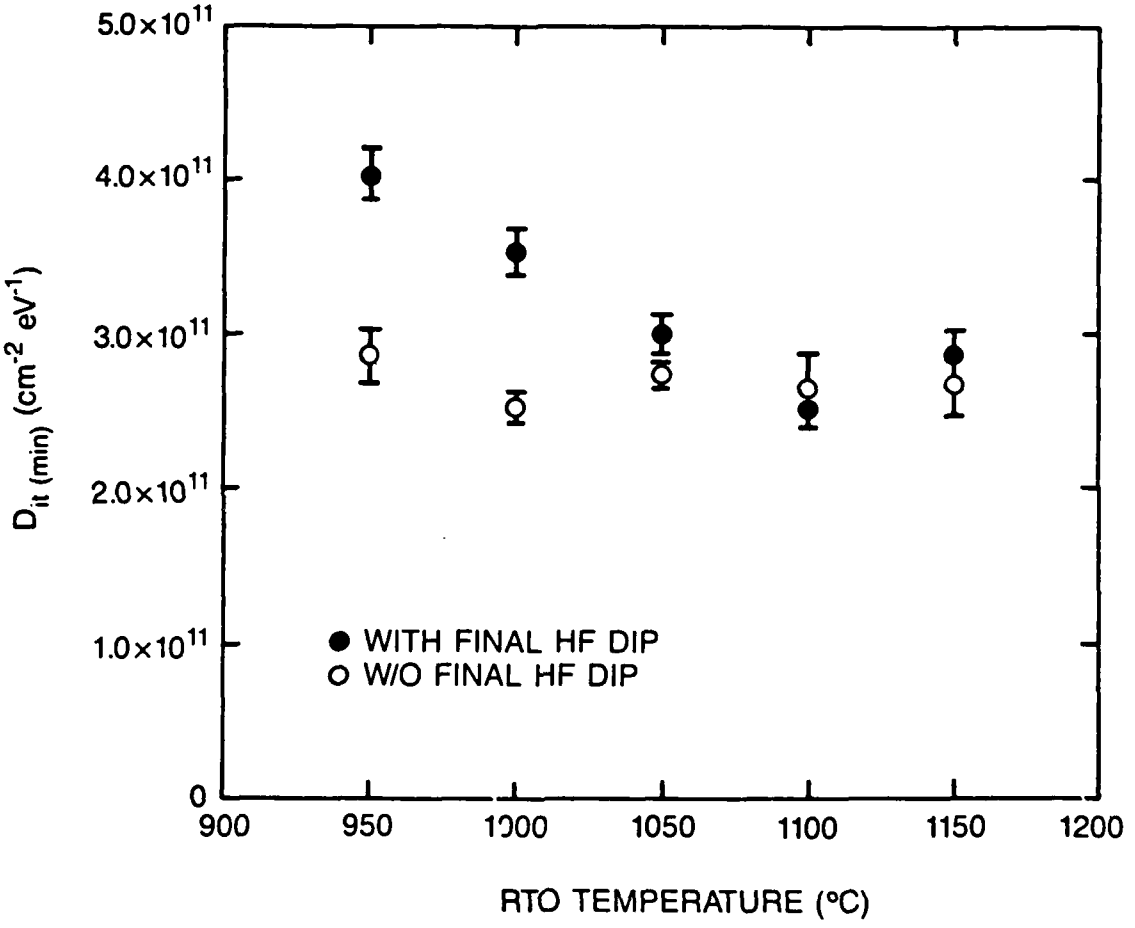


Figure 5.5.6: (a) FLATBAND VOLTAGE, (b) PRODUCT OF FLATBAND VOLTAGE AND GATE OXIDE CAPACITANCE, AND (c) MINIMUM SURFACE-STATE DENSITY VS RTO TEMPERATURE IN MOS DEVICES WITH TWO DIFFERENT PREOXIDATION CLEANINGS.


5.0×10^11
4.0×10^11
3.0×10^11
2.0×10^11
1.0×10^11
0
D_it (min) (cm^-2 eV^-1)
● WITH FINAL HF DIP
○ W/O FINAL HF DIP
900
950
1000
1050
1100
1150
1200
RTO TEMPERATURE (°C)

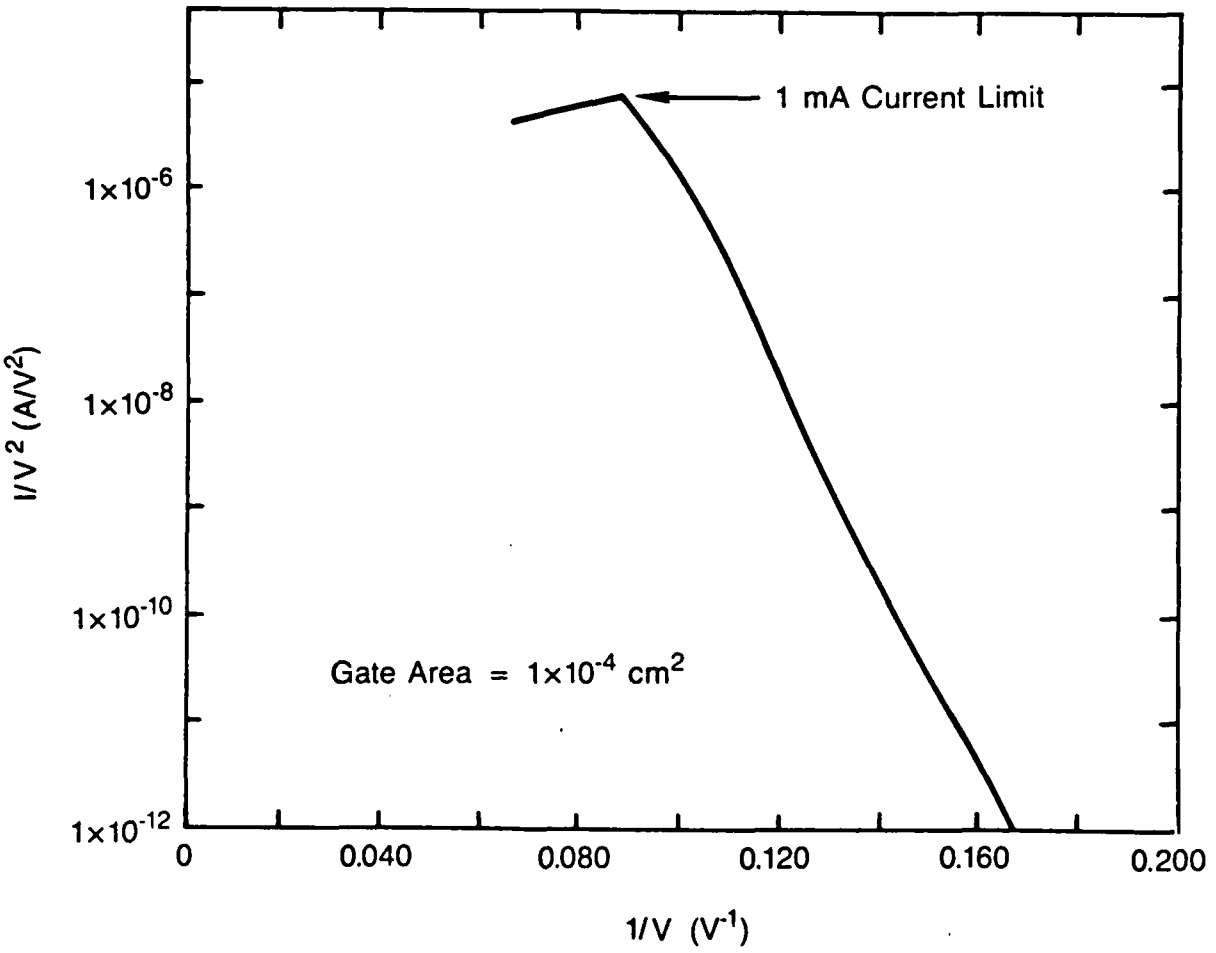


Figure 5.5.7: FOWLER-NORDHEIM PLOT OF GATE CONDUCTION in MOS devices with 79 Å gate oxide grown at 1050°C on wafers with the final preoxidation HF dip (gate area=100 μm×100 μm).

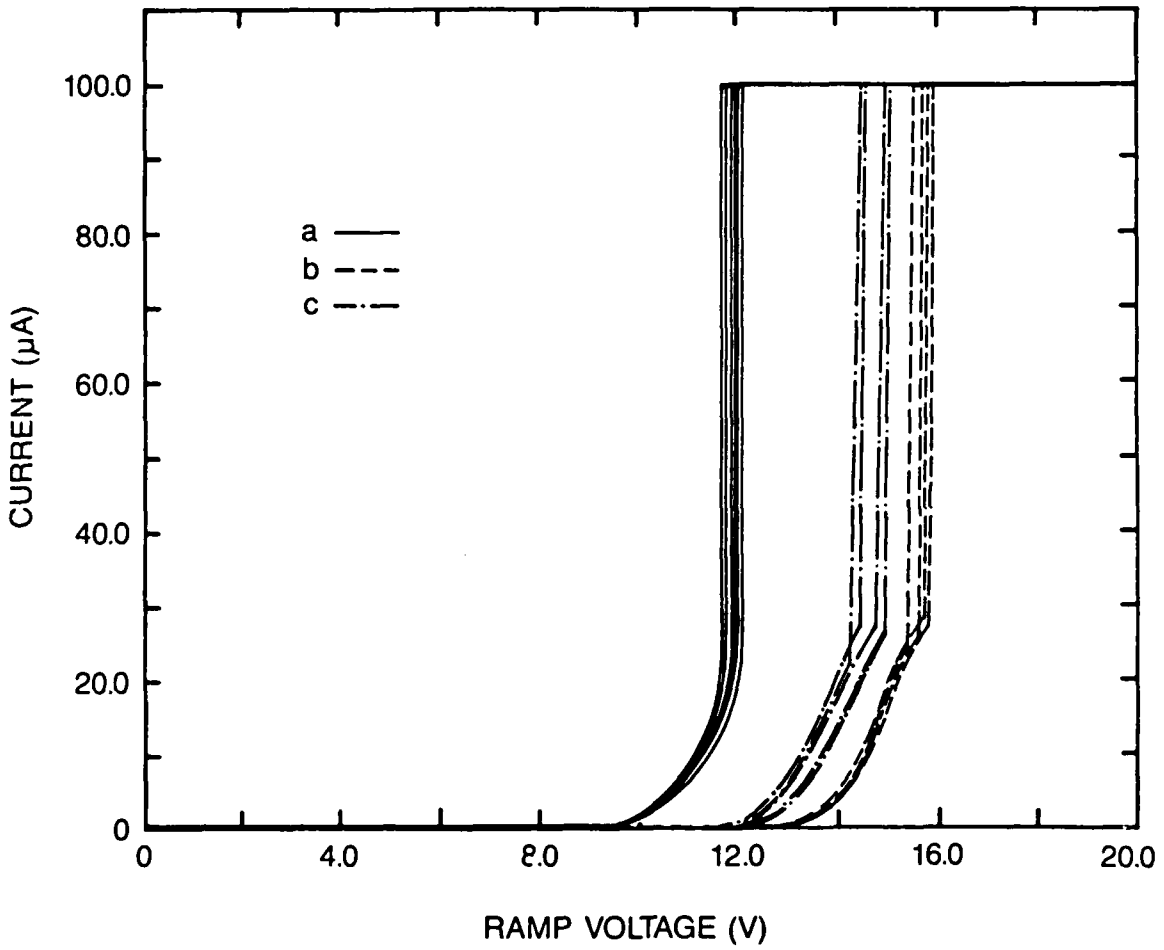


Figure 5.5.8: RAMPED-VOLTAGE BREAKDOWN CHARACTERISTICS OF MOS DEVICES with (a) 79 Å oxide grown at 1050°C with preoxidation HF dip, (b) 105 Å oxide grown at 950°C with preoxidation HF dip, and (c) 98 Å oxide grown at 950°C without preoxidation HF dip. Each set of curves contains multiple measurements on 10 $\mu m \times 10$ μm devices located on different test chips using a voltage ramp of 1 V/sec.

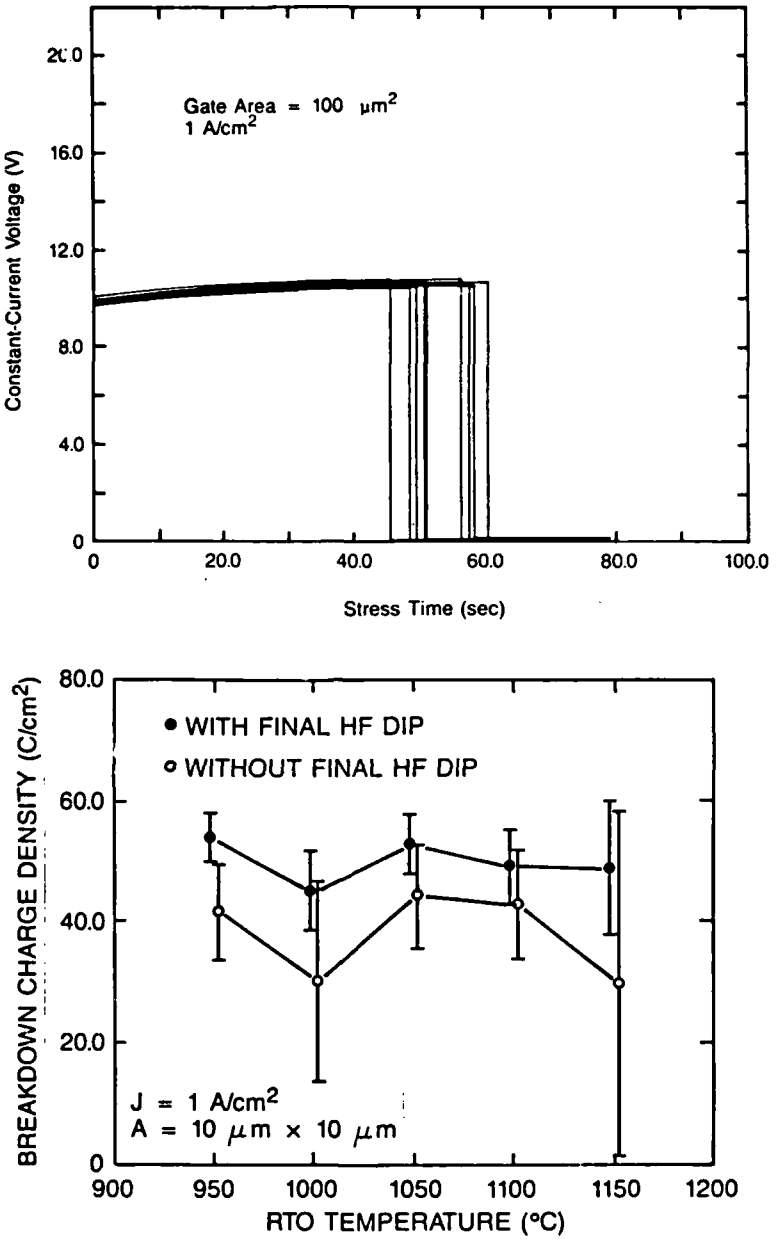


Figure 5.5.9: (a) CONSTANT-CURRENT STRESS TDDB MEASUREMENTS at a stress current density of 1 A/cm^2 in 10 μm×10 μm MOS devices with 79 Å oxide grown at 1050°C with preoxidation HF dip. (b) BREAKDOWN CHARGE DENSITY VS RTO TEMPERATURE at a stress current density of 1 A/cm^2 in 10 μm×10 μm MOS devices with and without the preoxidation HF dip.

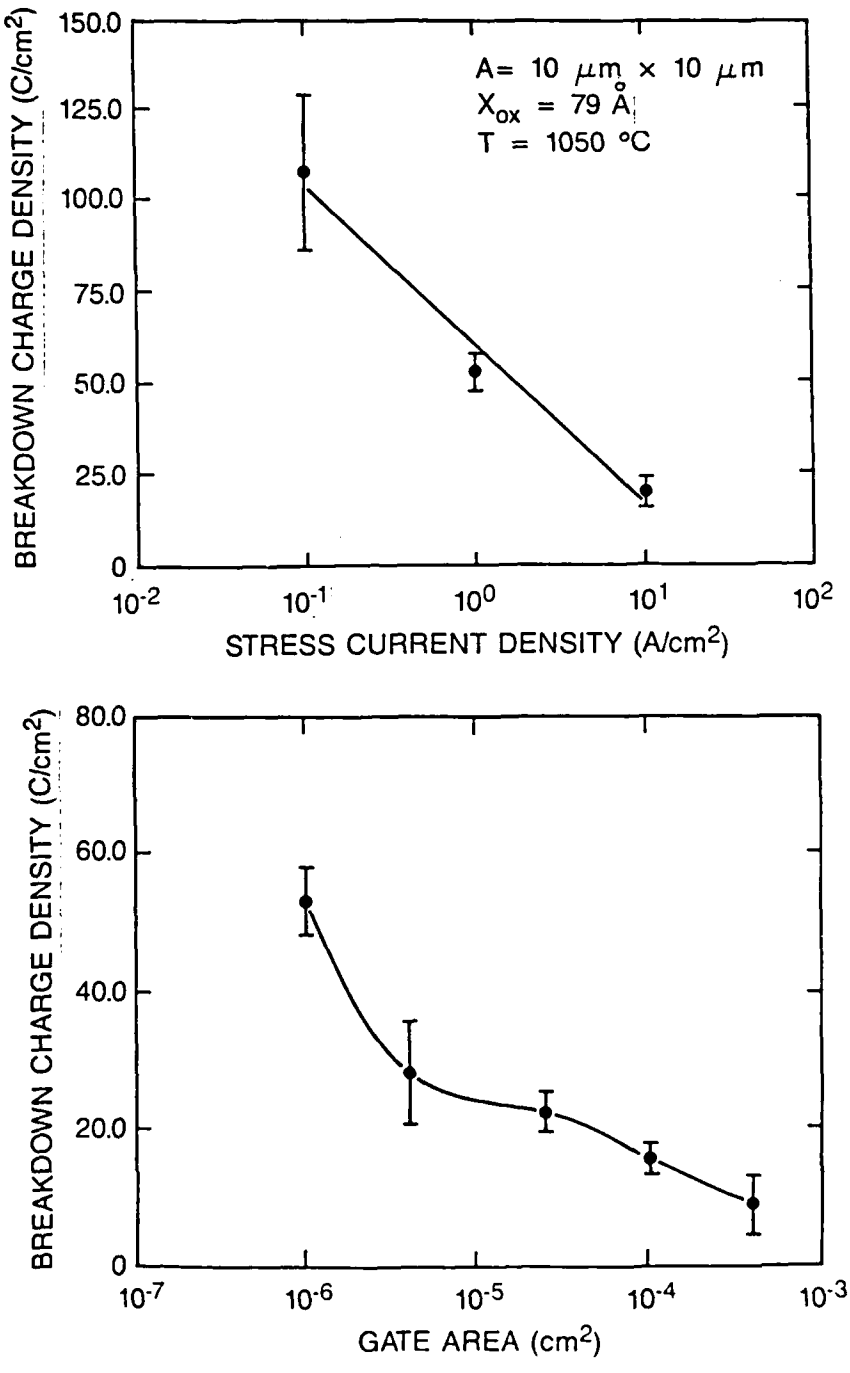


Figure 5.5.10: DEPENDENCE OF BREAKDOWN CHARGE DENSITY (Q_{bd}) ON (a) STRESS CURRENT DENSITY in 10 μm×10 μm devices (the solid line represents least-square fit of the measured data to the empirical equation), AND (b) ON THE GATE AREA (at 1 A/cm^2), in MOS devices with 79 Å oxide grown at 1050°C with preoxidation HF dip.

5.6 Template-Set Approach to VLSI Pattern Inspection

Soo-Ik Chae, James T. Walker, David H. Dameron, Chong-Cheng Fu, James D. Meindl

5.6.1 Abstract

A new approach is described for the automatic detection of defects in VLSI circuit patterns such as photomasks and wafers. It is based on morphological feature extraction using templates that represent a set of local pixel configurations within a specified window. These templates are stored in content-addressable memories (CAMs) to facilitate parallel comparisons of window-pattern scanning over a tested image. Maskable CAMs reduce the size of a template set substantially. Two error-detection algorithms are implemented to detect both random defects and dimensional errors.

5.6.2 Introduction

As the minimum feature size of VLSI patterns is reduced and the number of mask levels is increased, the inspection of masks and wafers becomes increasingly important to maintain a high yield. The inspection methods based on binary image patterns have been proposed[5.6.1] and are classified as *reference methods* for pixel-by-pixel comparisons and *local-property methods* based on local-pattern configurations. The local-property method does not require computer aided design (CAD) data as a reference or rigid alignment because every decision is based only on a local pattern; however, some patterns that do not violate the design rules, such as missing or additional patterns, cannot be detected. These errors are rarely caused by the patterning process itself and can be eliminated by reliable CAD and pattern-generation tools.

The VLSI patterns on the masks and wafers are normally restricted. Usually their edges are straight, and their directions are locked on a small set of angles such as {0, 90, 180, 270} in the Manhattan-style layout. In addition, there are design rules that limit the degrees of freedom of the patterns. Although acquisition of the pattern is another difficult problem, a binary image of the circuit patterns is assumed to be available for real-time inspection. This paper focuses only on image analysis.

An objective of pattern inspection is to locate the center of every window that contains an unacceptable pattern. This can be achieved by comparing a windowed image to a set of templates that represent the morphological properties of the circuit patterns. This template-set approach to pattern inspection of a binary image consists of the following steps:

Step 1: collect all acceptable or unacceptable local patterns as a template set

Step 2: obtain a local windowed image centered at a pixel

Step 3: determine whether any template matches the window pattern

Step 4: repeat steps 2 and 3 until all the pixels in the inspected image are scanned

Because the patterns are quite restricted, the acceptable or unacceptable patterns are well-defined. All possible acceptable patterns or their complement must be stored as a template set. There are several difficulties in generating the template set,[5.6.2] however, and a suitable approach must satisfy several conditions.

- **Reasonable size of the template set.** The template set is stored as a look-up table in a memory and, as a result, the number of templates must be reasonably small because the size of the memory is limited so as to be manageable.

- **Completeness of the template set.** A template set for an algorithm must include all templates that match the window patterns determined by the algorithm.

- **Fast searching.** Inspection throughput is determined by the speed of searching for any template matched to the window pattern.

- **High defect coverage.** The two types of defects that must be detected in VLSI pattern inspection are random and dimensional errors. The defect coverage of an inspection algorithm depends on the size of the window. It is difficult to detect both types of errors with a reasonably small window.

Goto[5.6.3] and Jarvis[5.6.4] published papers describing this template-set approach. Goto used a 3×3 window to lessen the template number. Each window pattern was classified into one of 16 predetermined categories and represented by a 4-bit word. Because the window was small, its capability of defect detection was poor. As a result, a hierarchical method based on a 5×5 array of 4-bit words was employed to detect a defect larger than the window. Jarvis applied this approach as a filtering step because it was difficult to obtain a small complete template set. Although the template set was not complete, it could filter out most of the acceptable window patterns so that the number of patterns to be further analyzed was minimized.

5.6.3 Solution

A solution to the above problems is proposed for easy implementation in VLSI circuits. An objective of the inspection is to detect unacceptable deviations from the intended patterns. As a measure of deviation, *edge roughness* must be defined so that random error can be determined precisely, and this will require additional definitions. For a discrete image I in a square grid system, $I(i,j)$ is a square picture element (pixel) where (i,j) represents its position. A pixel with the value of 0 (1) is called a 0 (1). A *4-neighbor* of a pixel $I(i,j)$ is any pixel $I(m,n)$ such that $|m-i|+|n-j|=1$. An *8-neighbor* of a pixel $I(i,j)$ is any $I(m,n)$ such that $\max\{|m-i|,|n-j|\}=1$. A *connected* pattern is a set of 1s such that each has at least a 4-neighbor equal to 1. An *edge pixel* is any 1 with at least a 4-neighbor equal to 0. A *contraction* operation C to a window pattern P sets every pixel p to 0 such that p is 1 and has at least an 8-neighbor in the window equal to 0. The pattern resulting from the application of k contraction operations to the window pattern P is denoted by C^kP. For window patterns A and B, $A \subset B$ if $B(i,j)=1$ for all i,j such that $A(i,j)=1$ and at least one pixel position exists such that $B(i,j)=1$ and $A(i,j)=0$. A *noise-free window pattern* has a straight edge with zero or one corner in the window. A pattern with an *edge roughness* of k is defined as any P such that P is a connected pattern, and a noise-free window pattern R exists such that $C^kR \subset P \subset R$, $P \not\subset CR$, and $C^{k-1}R \not\subset P$. It must be noted that this edge roughness is not a global but a local concept that is applied to window patterns. Pixel size is assumed to be small enough that patterns with an edge roughness of 1 are acceptable. Random-error patterns are thereby defined as patterns with an edge roughness of more than 1. This solution has three significant advantages.

- **Bit maskability.** The number of templates can be reduced substantially by changing every edge pixel of the template pattern into "don't-care." The number of required templates with "don't-cares" increases polynomially as the window size increases linearly, but it rises exponentially without "don't-cares." The completeness of the template set can be achieved easily because its size is reasonably small. In addition, the false error detection rate is reduced. As a result of quantization error, some degree of edge roughness may occur in the discrete binary image; however all the patterns with an edge roughness of 1 must be accepted. False error detection can be avoided by masking all pixel positions corresponding to edge pixels of the intended patterns during the comparison operation. This is an effective approach to the quantization-error problem.

- **Content addressable memory.** The speed of the matching operation can be independent of the size of the template set if the templates are stored in a content-addressable memory. Bit masking can also be implemented for "don't-care" conditions.

- **Multiple algorithms.** Two algorithms for *random error detection* (RED) and *dimensional error detection* (DED) are used to enhance the defect coverage in VLSI pattern inspection. There is no speed penalty because both algorithms can be processed in parallel.

This solution can be implemented in hardware for real-time inspection at the video signal rate because of its high speed and parallelism. The following steps are required to apply this approach to a specific problem.

- Determine the type of acceptable patterns.

- Develop error-detection algorithms and select a window for each.

- Construct a template set for each algorithm.

The window-pattern space is an N-dimensional binary space $\{0,1\}^N$ where N is the number of bits in the window. After the acceptable patterns are defined, the inspection algorithm to be used can be determined. A window for this algorithm is then chosen, and patterns are mapped into the window-pattern space as illustrated in Figure 1. There are two sets in the window-pattern space – one for acceptable and the other for unacceptable window patterns; either one can be selected as a template set.

5.6.4 Error-Detection Algorithms

In VLSI pattern inspection, the two types of defects of interest are random defects and dimensional errors. High defect coverage can be obtained by two detection algorithms; one is a RED algorithm that can detect irregular shapes and the other is a DED algorithm that can detect patterns with widths or gaps smaller than their minimum allowable values.

A RED algorithm is represented by the statement, *random-defect patterns have an edge roughness of greater than 1.* More precisely, however, some of the unacceptable patterns are characterized by edges that are not oriented in any allowable direction. In this algorithm, a square window is suitable for the extraction of the edge profile of the patterns. It should be large enough to detect most random features but must contain only one corner or less of the acceptable pattern. If two

corners are permitted, the number of templates in the set will increase by approximately an order of magnitude.

All acceptable window patterns are included in the RED template set because the subset in the window space corresponding to the acceptable patterns is smaller and easier to generate. The layout rules must be restricted so that the template set is simpler to generate. Assume that the layout is based on the Mead-Conway design rules[5.6.5] and all corners are on the λ grid points. It is also assumed that the minimum feature size equal to 2λ is 10 pixels. The maximum size of the window is then limited to 5 pixels in each side so that the window permits only one corner.

The template patterns for the RED algorithm are local *two dimensional* (2-D) patterns because its defect coverage is superior to that obtained from any other local-property method that does not fully use the 2-D information. The RED algorithm for the 1λ square window(Figure 2a) can be refined by the corner number in the window, edge roughness, and edge direction and can be stated as follows.

> *If there is one corner or less in the window and the direction of any edge is one of the locked directions and the roughness of any edge is less than or equal to 1, the window pattern is acceptable.*

A pattern can be acceptable in the RED window but unacceptable in the DED window(Figure 2b) because the minimum gap and width requirement is violated. The DED algorithm is as follows.

> *There is an error pattern in the window if (1) there is a pair of parallel edges within a circle or if (2) some pixels within the circle differ in value from those on the circle when all pixels on the circle have the same value. Here, the diameter of the circle is equal to the minimum feature size.*

Because a square grid system is employed, an octagon must be used as an approximated circle. All pixels inside the octagon need not be in the window because of redundancy. All pixels in the boundary of the octagon (condition 1) and several additional pixels inside the octagon (condition 2) must be included in the window, as illustrated in Figure 2b. The unacceptable patterns are selected as the template set for the DED algorithm because it can be represented by a smaller number of templates.

If the minimum allowable width and gap differ, separate template sets and windows are needed for gap and width error detections, respectively; for simplicity, it is assumed that they are equal.

In Figure 3, the subsets of patterns whose window patterns match the template sets from RED and DED are labeled A and B, respectively. The set of acceptable patterns is $A - B$ because the RED template is for acceptable window patterns and the DED template is for window patterns with errors. The DED algorithm can be adjusted to eliminate all unacceptable patterns from the subset corresponding to the RED template set.

5.6.5 Template Set

For a window w with N bits, a *template* is defined as a subset of the window space $\{0,1\}^N$ and is represented as a doublet with two N-bit words $\{t_p, t_m\}$ - a *pattern word* t_p and a *mask word* t_m. If $w_k = t_{pk}$ for all k such that $t_{mk} = 0$, w is *matched* to the template $\{t_p, t_m\}$, where

$t_p = (t_{p1}, t_{p2}, ..., t_{pN})$, $t_m = (t_{m1}, t_{m2}, ..., t_{pN})$, $w = (w_1, w_2, ..., w_N)$. If $t_{mi} = 1$, then t_{pi} is "don't-care" and w_i will be matched to t_{pi}. Actually, a template can be implemented with a maskable CAM. If w is matched to at least one template in the template set T, T is said to match w. If T matches only every acceptable window pattern, it is complete for acceptable patterns; if it matches at least one window pattern during the scanning only for each error pattern, it is complete for unacceptable patterns.

Template-Set Generation

A template set is optimal when it is small in size and maximum in defect coverage. The template set can be generated by the empirical collection method, primitive pattern method, and algorithmic property method.

- **Empirical collection method.** Window patterns can be collected as templates by scanning a defect-free image until there are no more templates. It is an empirical extraction process of local patterns without direct use of an error-detection algorithm; however, it is simple and straightforward. This method encounters two problems. First, it is difficult to ensure the completeness of the template set because it is not easy to scan all possible patterns and, second, it is difficult to determine when to finish collecting the templates.
- **Primitive pattern method.** Because VLSI circuit patterns can be represented by a set of primitive patterns, real images are not necessary to generate the template set. In the Manhattan style, the primitive patterns are a square pattern with each side larger than the minimum feature size and its complementary pattern, which is a square hole. It is basically the same as the empirical method but extraction from only primitive patterns is required to guarantee the completeness of the template set.
- **Algorithmic property method.** Template generation can be simplified by using the local property of the patterns in the image and the property of the algorithm itself. It is an effective method if the patterns required to match templates are simple in the error detection algorithm.

RED Template Set

It is easiest to construct a RED template set for acceptable patterns by using the primitive pattern method because the primitives are small in number. A template-generation algorithm, as outlined below, obtains window patterns from an image that includes only primitive patterns and changes each of its edge pixels to "don't-care." It guarantees the completeness of the template set.

```
begin(* template set generation *)

    set templateset empty;
    repeat
read a new pixel data;
form a new window pattern W;
```

```
T = W;
for each pixel W(i,j) do begin
    if W(i,j) is an edge pixel then begin
T(i,j) = don't care;
    end else begin
T(i,j) = W(i,j);
    end;
end;
if T is not in templateset then begin
    templateset = templateset + T;
end;
   until all pixel data are scanned;

end;(* template set generation *)
```

As an example, a window pattern and its matched templates are illustrated in Figure 4.

DED Template Set

The DED template set is constructed only for unacceptable patterns and its purpose is to detect the errors that cannot be distinguished by the RED algorithm. All irregular patterns can be detected by the RED algorithm; however, unacceptable rectangular patterns that violate the minimum width and gap requirement can also be matched to a RED template. These errors can be observed by creating a DED template that corresponds to every possible rectangular error pattern positioned at the center of the DED window such that it must be matched to the template at least once during scanning. The algorithmic property method is suitable for template-set generation because rectangular patterns are simple. An error pattern and a template matched to it are illustrated in Figure 5.

5.6.6 Analysis and Simulation of Defect Coverage

Defect coverage is perfect if there is neither a missed defect nor a false detection. The probability of false detection caused by noise is assumed to be negligible. To evaluate defect coverage with two RED and DED templates sets, it can be verified by simulation but coverage analysis is more effective. It is assumed that the RED square window2 is N pixels in each side and the DED octagon window is $2N + 1$ in diameter where the minimum feature size is $2N$ pixels. The following type of defects can be completely detected:

- **Every pattern with an edge roughness of 2 or more.** All patterns except those with an edge roughness of 1 or 0 can be detected as errors by the RED algorithm because a RED template set includes all possible templates for patterns with an edge roughness of 1 or 0.
- **Every rectangular pattern with a minimum width or gap of less than N – 1 pixels.** It can be detected by the RED template set because it has two parallel edges in the window; however, there is no template for the window pattern with two parallel edges in the RED template set.

- **Every rectangular pattern with a minimum width or gap of less than 2N pixels and greater than N – 2 pixels.** It can be detected by the DED template set. If its maximum width or gap is greater than or equal to $2N$, it is detected by condition (1) of the DED algorithm; otherwise, it is detected by condition (2).

In summary, every pattern with a random shape and every rectangular pattern with a width or gap of less than $2N$ pixels can be detected by the RED and DED template sets. Undesired additional rectangular patterns whose minimum widths are greater than $2N$ cannot be detected. The probability of such error patterns, however, is negligible.

The algorithms have been simulated to verify their correctness and the defect coverage, and the results were also used to determine the number of detections for a particular error. Figure 6 is a binary SEM image of contact-hole patterns in photoresist. The RED templates were generated to tolerate the rounding effect at each corner, and the DED templates were generated to detect only feature sizes that were less than 80 percent of the allowable minimum size. These constraints resulted in an RED set with 128 templates and a DED set with 16 templates. A simulation result with these template sets is shown in Figure 6. The positions where error patterns exist are marked with a star for dimensional error and a small circle for random error. For convenience the inspection operation was not applied in the region outside of dotted line. The two types of errors in this image were the unintended bridging and narrow gap between contact holes caused by overdevelopment of resist, and both were correctly detected.

The results of RED template generation for various sizes of square windows in the Manhattan-style patterns are summarized in Table 1. As the window increases, the memory required to store the templates grows rapidly, which limits the size of the window.

5.6.7 Custom VLSI Circuits

As illustrated in Figure 7, two functional blocks for windowing and comparing are required to inspect a serial binary input as a tested image. The windowing block forms a 2-D window pattern from a serial video signal and is basically a set of shift registers. Its output is connected to the bit-line drivers that drive the bit lines of the CAM array of the comparing block.

The matching block consisting of maskable CAM cells compares the window pattern to all templates stored in the content-addressable memories in parallel, and N maskable CAM cells are tied together to a match line where N is the number of bits in a window pattern. Because there is a match line for each template, the match signal is the output of an OR operation on all match lines. A maskable CAM cell with 17 transistors (Figure 8) can store one bit for the pattern word and one for the mask word. Two custom ICs have been designed and are being fabricated. A RED IC can store 128 templates for the 25-bit window and a DED IC can contain 64 templates for the 32-bit window.

Figure 9 is a block diagram of a prototype inspection system. The input is a serial binary image of the VLSI circuit patterns, and the output is a color display. The delays are used to locate the centers of the RED and DED windows at the same pixel and to synchronize the outputs of the RED and DED chips to the delayed input signal, which corresponds to the center pixel of the RED and DED windows.

5.6.8 Conclusion

The template-set approach is a simple and effective method for resolving the problem of VLSI circuit pattern inspection. A real-time inspection can be accomplished with fast content-addressable memories. The advantages of this approach are summarized as follows.

- "Don't-care" is employed to reduce the number of templates and to resolve the problem of quantization error effectively.
- The maskable content-addressable memory is used as a storage for the template set and parallel-comparison unit.
- The content-addressable memory required to store the template sets is orders of magnitudes smaller in volume than the CAD data of the patterns.
- The template set can be adjusted if the set of error patterns to be covered is changed.
- Multiple template sets can be employed to enhance defect coverage if a template set cannot detect all the error patterns.
- Real-time inspection is possible because of the high speed and parallelism of a CAM array.
- This approach is compatible with the raster-scanned image because the window pattern at every pixel position must be compared to the template set.

Because outputs of this inspection method are positions where the unacceptable window patterns exist, it can be used for analyzing the unacceptable window patterns to classify the type of defects, which is an area of further research.

5.6.9 References

[5.6.1] R. T. Chin and C. A. Harlow. Automated visual inspection : a survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 557–573, Nov 1982.

[5.6.2] M. Mese M. Ejiri, T. Uno and S. Ikeda. A process for detecting defects in complicated patterns. *Computer Graphics and Image Processing*, 1973.

[5.6.3] N. Goto and T. Kondo. An automatic inspection system for mask pattern. *Proc. 4th Intl. Joint Confon. Pattern Recognition*, 1978.

[5.6.4] J. F. Jarvis. A method for automating the visual inspection of printed wiring boards. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 77–82, Jan 1980.

[5.6.5] C. Mead and L. Conway. *An Introduction to VLSI Systems.* Addison Wesley, 1980.

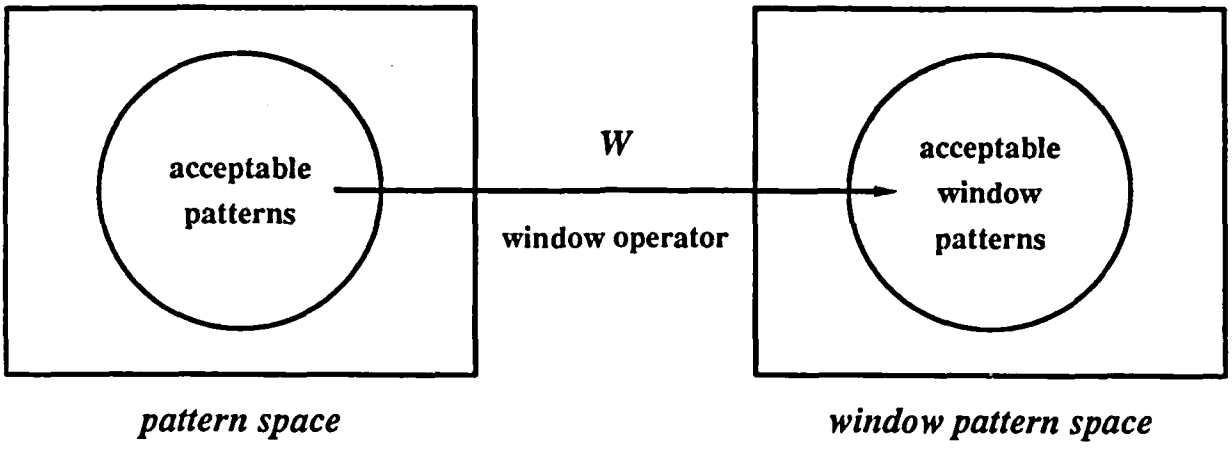


Figure 5.6.1: Mapping of patterns to the window-pattern space.

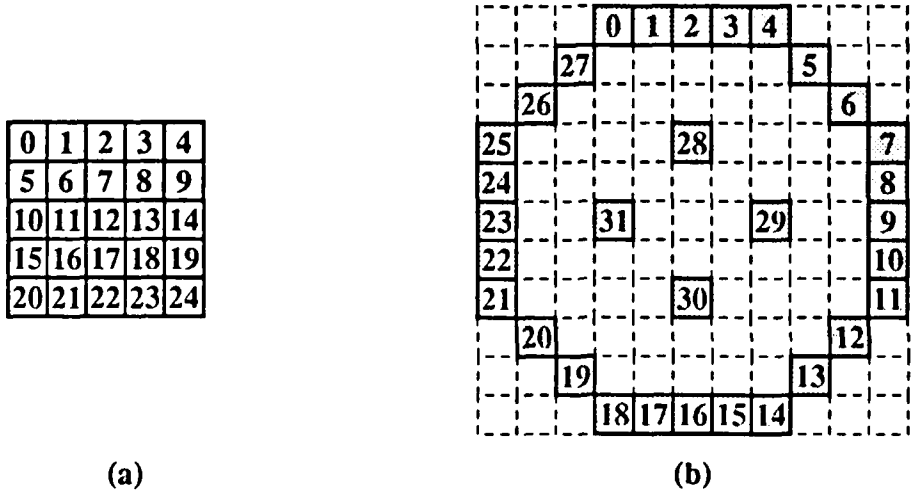


Figure 5.6.2: Window shape for DED and RED. (a) A 25-bit RED square window. (b) A 32-bit octagon DED window.

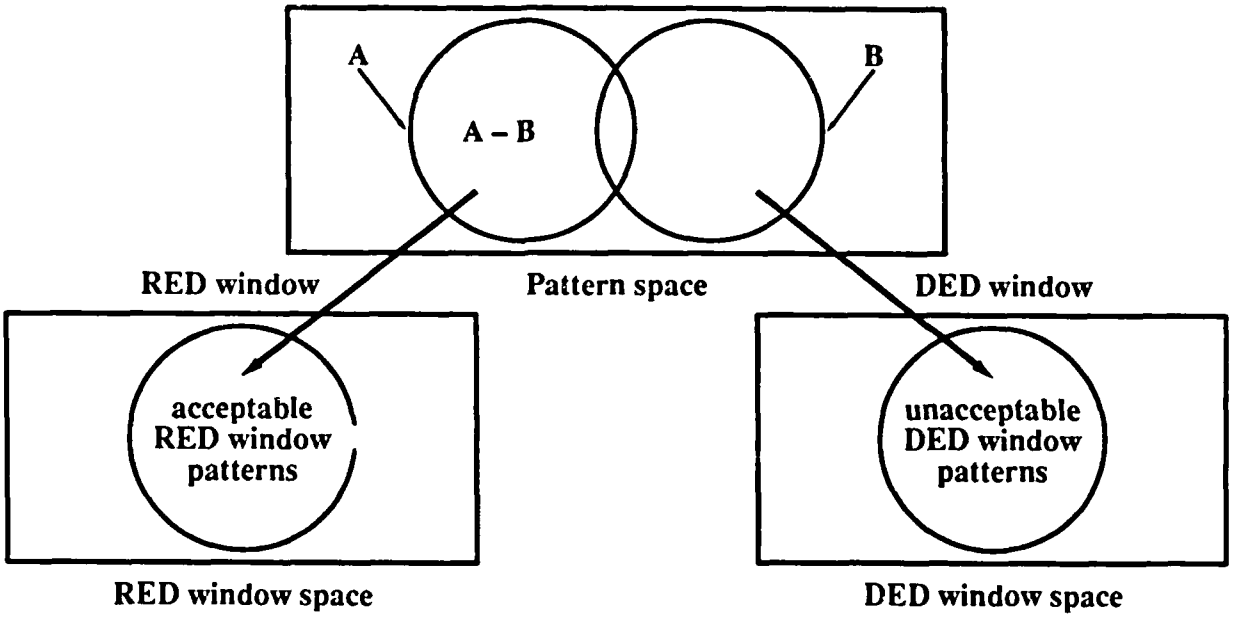


Figure 5.6.3: Acceptable pattern sets for the RED/DED algorithms.

Table 5.6.1: Number of RED templates versus size of window.

Window dimension N	3 × 3	4 × 4	5 × 5	6 × 6	7 × 7
Window size (bits) N^2	9	16	25	36	49
Template size (bits) $2N^2$	18	32	50	72	98
Template number T	32	72	128	200	288
Memory required (bits) M	576	2304	6400	14400	29264

pixel number	0	24
pattern word	001110011100111111111111	
mask word	001000010000100110000000	

(a)

0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	23	24

(b)

Figure 5.6.4: Examples of the RED template and a window pattern. (a) RED window. (b) Window pattern matched to it.

pixel number	0	31
pattern word	00000000000000000000000000001011	
mask word	11111111000111111111110001110100	

			0	1	2	3	4			
		27						5		
	26								6	
25					28					7
24										8
23			31				29			9
22										10
21					30					11
	20								12	
		19						13		
			18	17	16	15	14			

(a) (b)

Figure 5.6.5: Examples of the DED template and an error pattern. (a) DED template. (b) Error pattern matched to it.

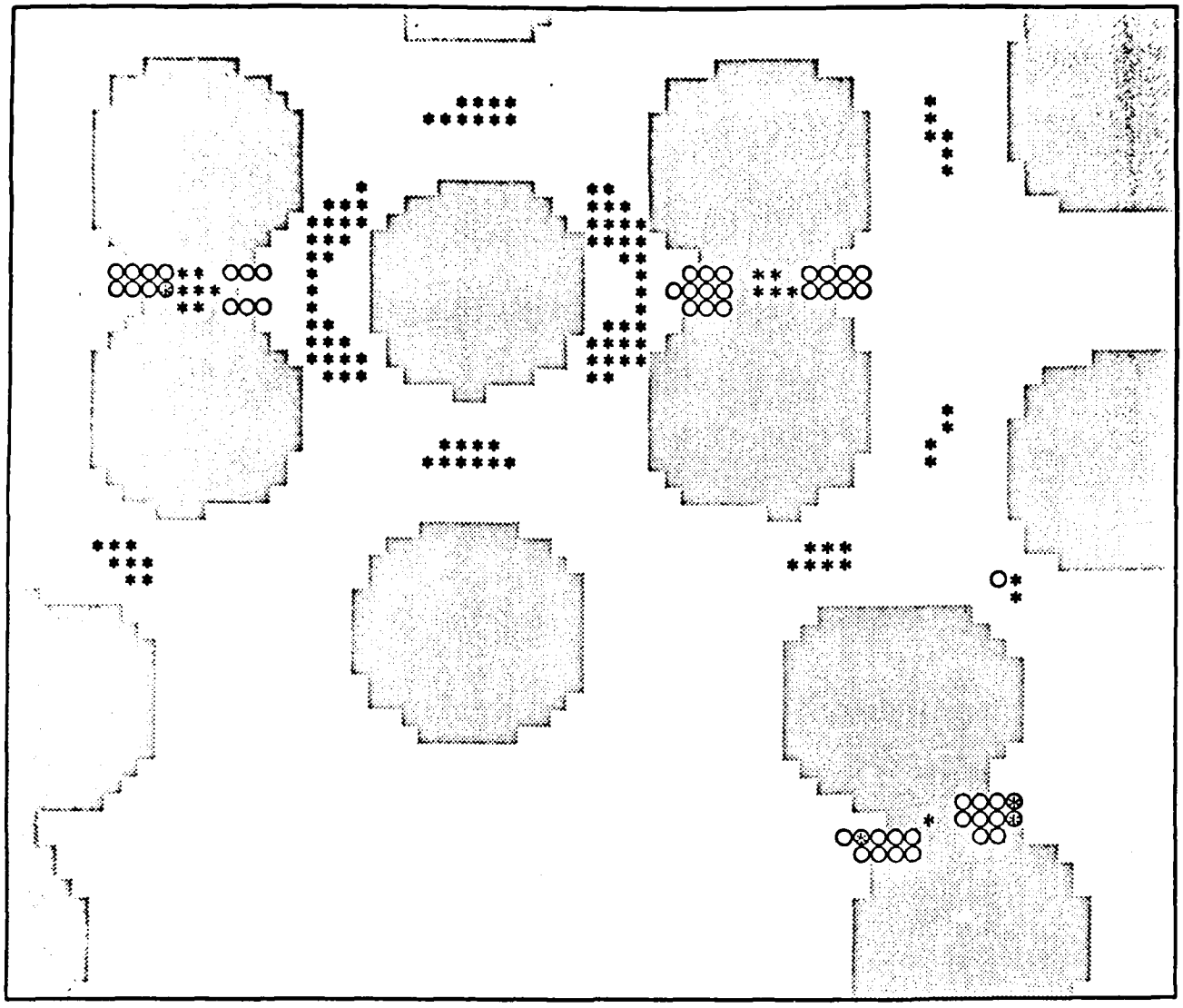


Figure 5.6.6 SEM image of contact-hole patterns in photoresist and simulation result. The star indicates DED, and the small circle denotes RED.

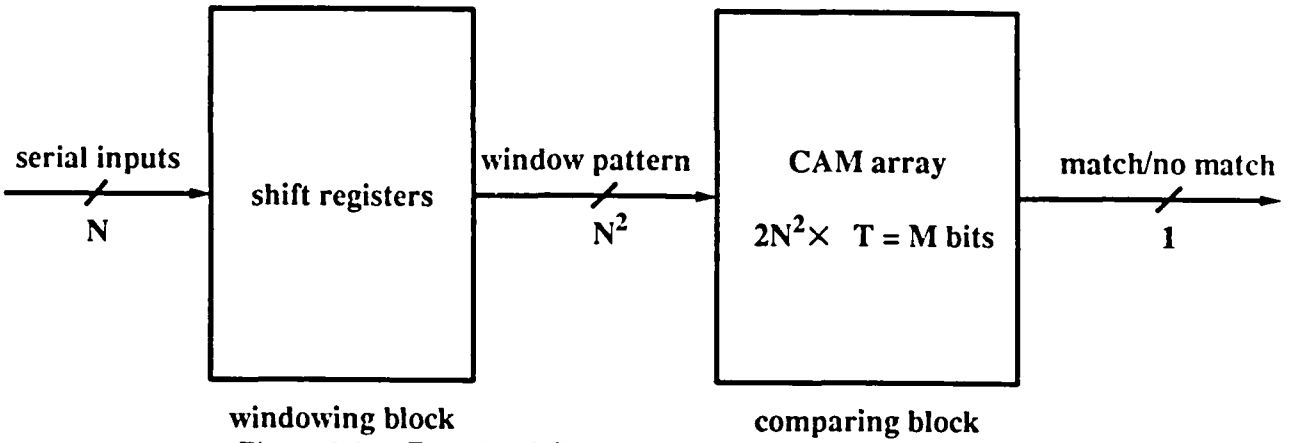


Figure 5.6.7: Functional diagram of an VLSI inspection circuit.

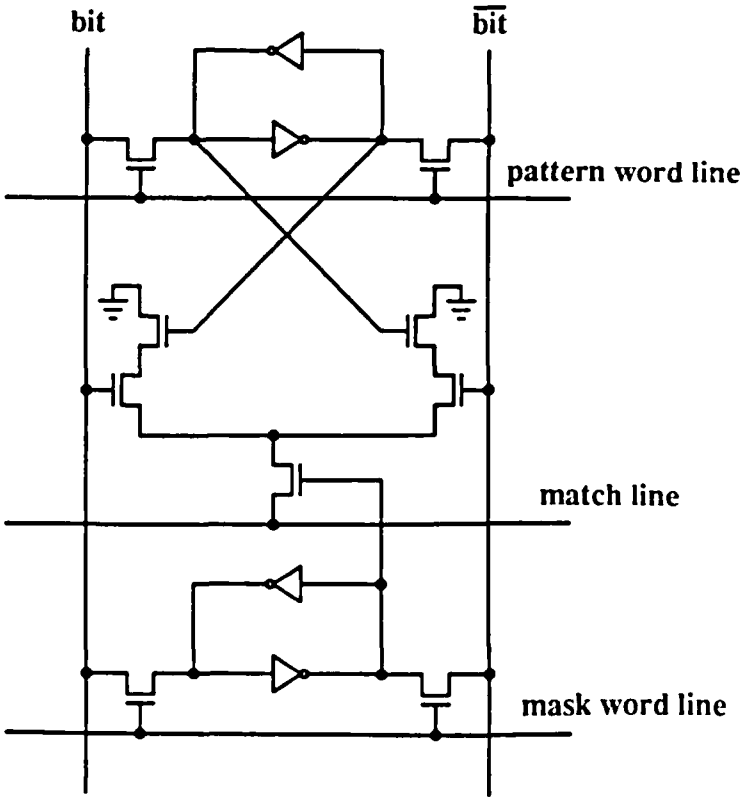


Figure 5.6.8: Circuit diagram of a maskable content-addressable memory cell.

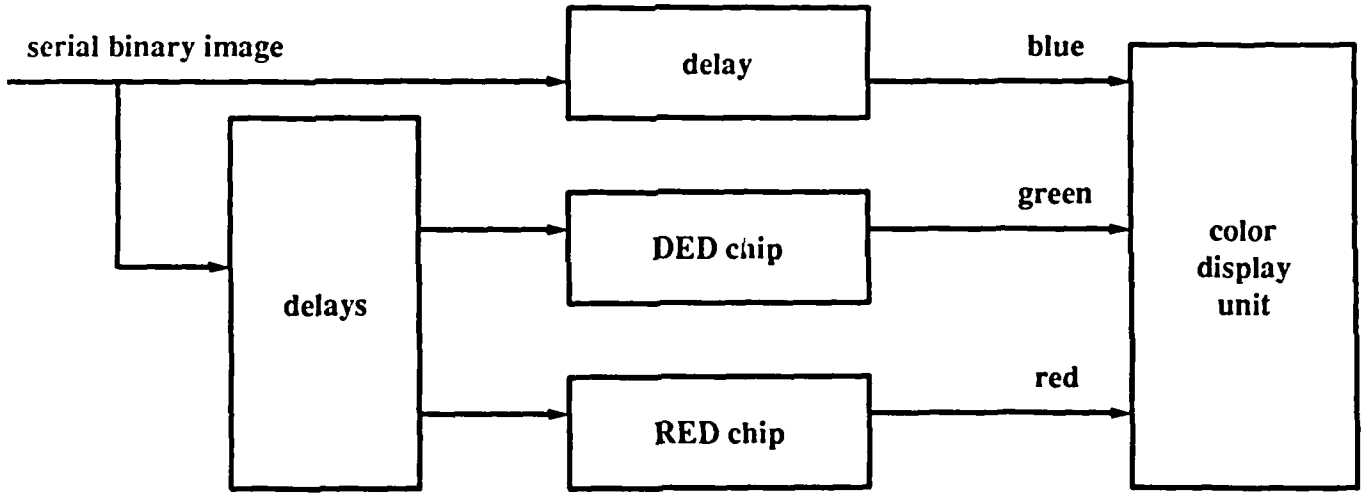


Figure 5.6.9: Block diagram of a prototype inspection system.

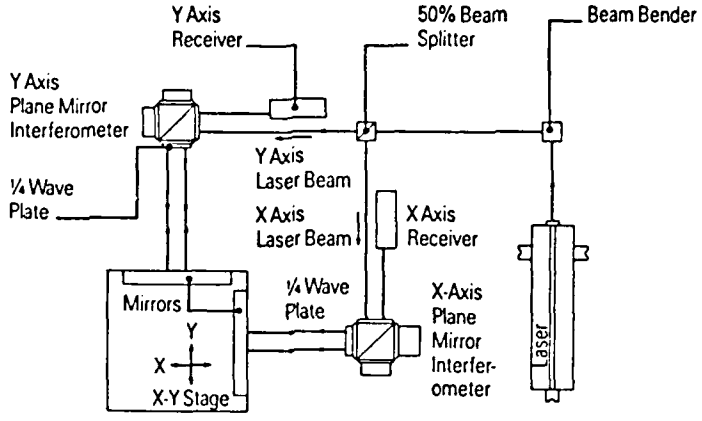


Figure 5.7.1: Laser interferometer system.

5.7 Characterization of Stage Drifts and Intrinsic Alignment Precision of the Ultratech Stepper

Chong-Cheng Fu

5.7.1 Introduction

The Ultratech 1× stepper has been the principal photolithgraphic tool for VLSI patterning in the Stanford Integrated Circuits Laboratory since 1983. In addition to high image resolution, it is capable of accurate overlay for feature sizes down to 1.25 μm. Such overlay accuracy, however, can be obtained only by using a *field-by-field* alignment strategy.

The stepper employs a Hewlett-Packard 5501A helium-neon Zeeman laser interferometer system to measure X and Y stage displacements. This system, as shown schematically in Fig. 5.7.1, is a two-axis plane mirror interferometer design, and is capable of resolving displacements on the order of 0.04 μm when properly protected from environmental fluctuations. In an effort to reduce the capital and operating costs of the stepper, however, Ultratech has chosen not to use an environmental control chamber. This cost effective move can lead to errors caused by temperature and humidity fluctuations. As a result, in a typical fabrication environment, it is very difficult to maintain a long-range spatially and long-term temporally stability in the measurement of stage position.

Such an environment-induced drift can be demonstrated by an experiment in which an alignment target fabricated on a wafer is repeatedly scanned by the target detection system of the stepper, and its position is measured by the interferometer each time. Ideally, the data should exhibit only small variations reflecting the intrinsic precision in the target detection mechanism, namely the repeatability of locating a target feature. As observed in Fig. 5.7.2, however, a substantial drift component usually dominates. In this example, drifts in excess of 1 μm occurred in both dimensions

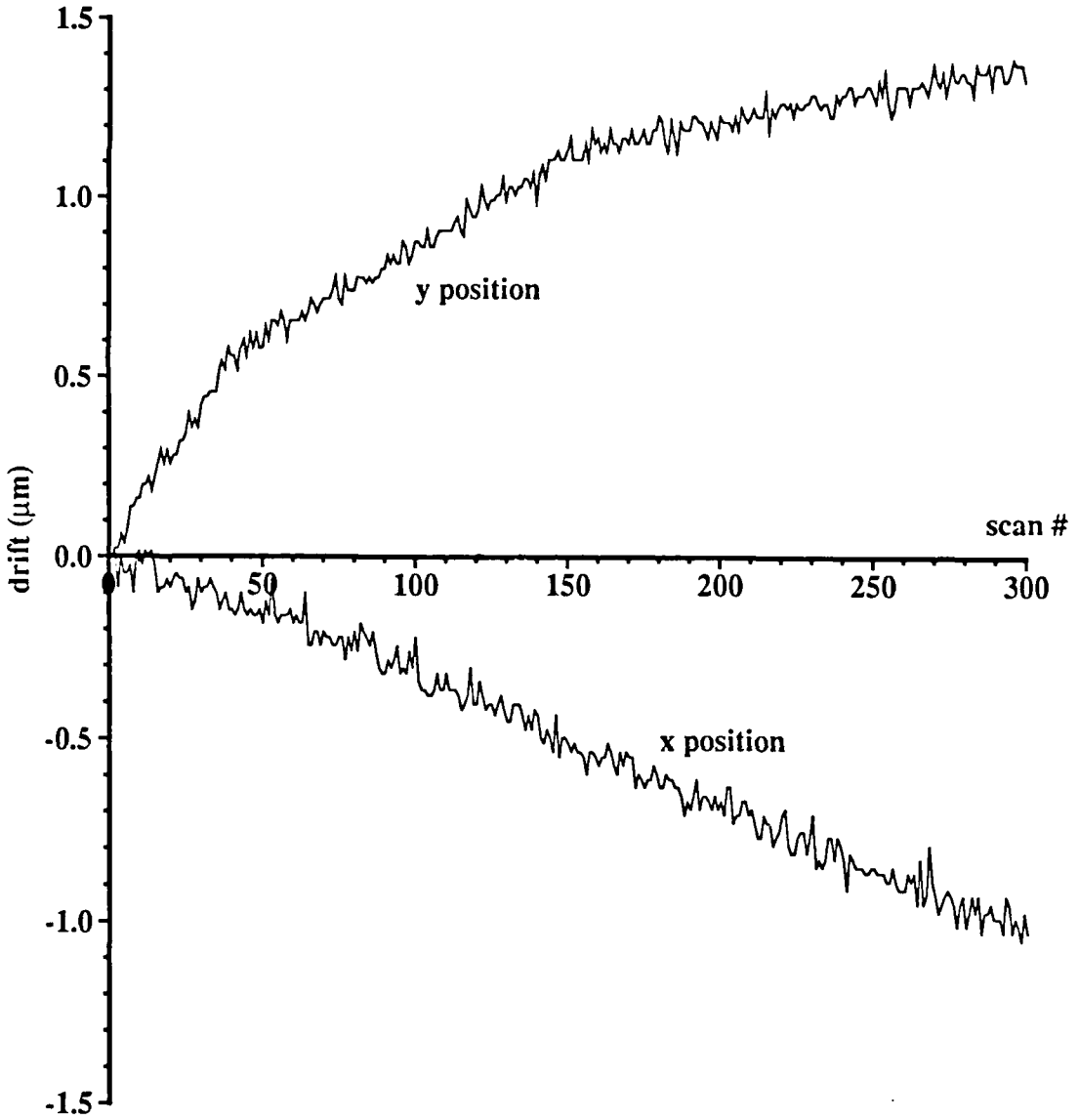


Figure 5.7.2: Measured target positions in 300 successive target scans, showing the environment-induced drift in the interferometer system.

during 300 repeated scans, which took approximately 10 minutes. It is therefore obvious that frequent local registrations are necessary for this stepper system to achieve acceptable overlay.

Another implication of this phenomenon is that intrinsic alignment precision of the stepper can not be evaluated without decoupling or cancelling the long-term drift, which itself varies substantially with the environmental variations.

5.7.2 Approach

The repetitive scanning scheme described above can be modified to enable quantitative evaluation of the intrinsic alignment error. Taking advantage of the short-term short-range stability of the laser interferometer system, two target scans on either the same target or two closely adjacent targets are performed in rapid succession, normally within two seconds. As suggested by Fig. 5.7.2, the drift during such a short period of time is negligible ($< 0.01\,\mu$m). In effect, the two scans provide each other with a constantly updated reference that tracks the drift, and the drift can therefore be cancelled out by subtracting one measurement from the other.

To statistically analyze this scheme, let the measured x positions from the first and the second scans be

$$x_1^* = x_1 + x_d \tag{5.7.1}$$

and

$$x_2^* = x_2 + x_d, \tag{5.7.2}$$

respectively, where x_1 and x_2 are the *true* scanned positions with respect to the system origin, and x_d is the drift component. Based on the discussion above, the drift can be assumed to be the same in both measurements. Defining $\delta = x_1^* - x_2^*$, then from Eqs. 5.7.1 and 5.7.2 $\delta = x_1 - x_2$. If both scans are performed on the same target, then x_1 and x_2 have the same distribution. Assuming further that x_1 and x_2 are *statistically independent*, i.e., the outcome of one scan is not affected by that of the other, it can be shown that $E\delta = 0$, and $\sigma_\delta^2 = 2\sigma_x^2$. Similar result can be obtained for the y position. Therefore, it is possible to experimentally estimate the intrinsic alignment precision even in the presence of a dominating stage drift.

5.7.3 Results and Discussions

The stepper was programmed to scan a single target feature at the highest repetition rate allowed by the system operation. Measured stage positions were sent to a VAX computer through a RS232 data link installed initially to enhance the stepper operation management capability for further analysis. The scans were performed with a non-actinic illumination which does not significantly affect the optical properties of the photoresist during the experiment.

As an example, the results from a 300-pair scan series can be summarized in Fig. 5.7.3, where histograms of the δ variables as defined above are plotted for x and y positions. As predicted, zero-mean distributions were obtained in both axes. Because the δ-variable overestimates the standard deviation by a factor of $\sqrt{2}$, the actual standard deviation values are $0.02\,\mu$m for both x and y scans.

Clearly, such variations are rather insignificant in comparison to the total RMS overlay precision (specified as $0.18\,\mu$m at one sigma[5.7.1]), which includes the contributions from lens-to-lens distortion, rotational error, reticle overlay, process variations, and so on. This is an important basis for implementing the *vote-taking lithography* scheme[5.7.2], in which the overlay among the participating fields is claimed to be better than the specification because the same target is visited by all fields.

5.7.4 Summary

By using a repetitive-scan scheme two important parameters, interferometer drift and intrinsic alignment precision, for the Ultratech stepper or other systems with similar design can be si-

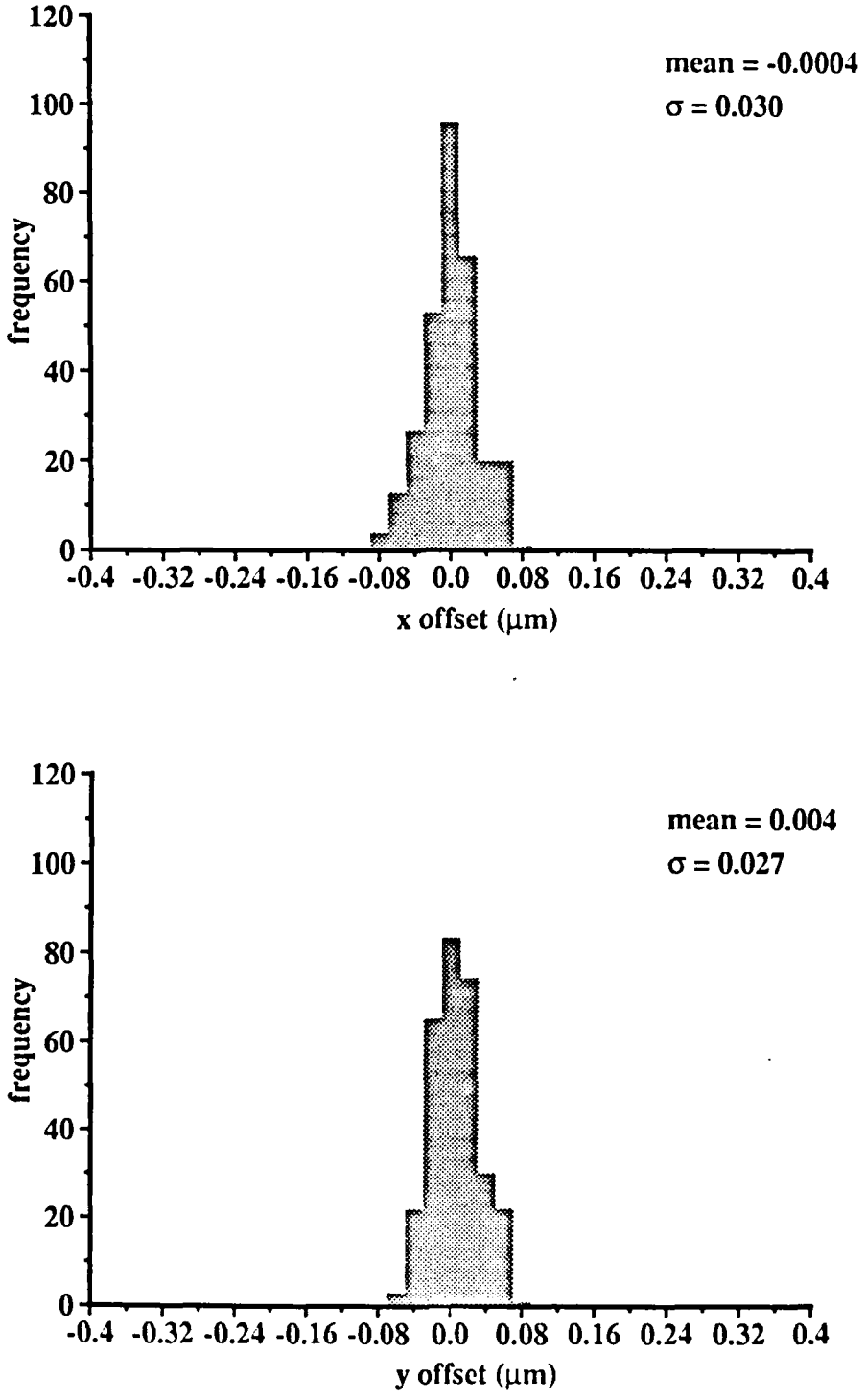


Figure 5.7.3: Histograms of the δ-variables obtained from 300 pairs of scans performed on a single resist target feature on silicon substrate.

multaneously characterized. The drift rate on our Ultratech stepper in the present laboratory environment is on the order of 1 μm in a 10-minute period. Its intrinsic alignment precision was found to be approximately 0.02 μm at one sigma, which represents only a very minor part of the total specified overlay precision. The data link between the stepper and the laboratory computer has made this task easy to perform.

5.7.5 References

[5.7.1] *UltraStep 1000 Product Description Guide.* Ultratech Stepper, Santa Clara, CA, 1984.

[5.7.2] C.C. Fu, D.H. Dameron and A. McCarthy. Elimination of mask-induced defects with vote-taking lithography. *Optical Microlithography V, SPIE J.*, 633, 1986.

5.8 Electron-beam Direct Write

Peter Wright, William Loh, David Dameron, Chong-Cheng Fu, and Krishna Saraswat

5.8.1 Introduction

Much of the increased performance for VLSI circuits has come from scaling the lateral device dimensions. As devices are scaled, however, parasitic aspects of a circuit such as contact resistance can degrade the gain and reduce the speed predicted by scaling laws. Knowledge of these circuit performance limiting factors is necessary in order to enhance speed for next generation ULSI circuits. Fabrication of these submicron ULSI devices and circuits requires a tool that can pattern small dimensions. Electron-beam lithography with a tri-level resist can be used to fabricate submicron devices to gain understanding of the next generation of ULSI circuits [5.8.1].

A single layer of an electron resist could be used for much of the processing of very small devices. But as the lateral dimensions of devices shrink, the thickness of the resist must also be scaled to obtain a higher resolution. Unfortunately there are problems associated with thin resist layers. Thinner resists tend to have inferior integrity, poor linewidth or critical dimension (CD) control over steps and other topography, and insufficient plasma etch resistance. Tri-level resist (TLR) is able to solve these problems at the expense of additional processing steps.

A TLR is made of three separate layers: a planarization layer, an isolation layer, and an imaging layer. The planarization layer is the bottom layer which smooths the substrate topography. The isolation layer separates the imaging and planarization layers, thereby preventing them from intermixing, and acts as a mask during the etching of the planarization layer. The imaging layer is an electron resist that is exposed with an electron-beam lithography system and developed similarly to photoresist.

Patterning the TLR is conceptually rather simple although many difficulties can be encountered during its development. The TLR is exposed and the imaging layer is developed. The isolation layer is then etched using the imaging layer as a mask, followed by an anisotropical etching of the planarization layer using the isolation layer as a mask. The final result is a mask for subsequent pattern transfer which has good CD's and is thick enough to resist plasma etching.

In this study, a TLR process was used to fabricate contact resistance test structures, to demonstrate the feasibility of this process, and to understand the electrical characteristics of very small contacts.

5.8.2 TLR Processing

A number of problems were encountered while developing the TLR process. Many of these problems and their solutions are discussed in this section.

The first material examined for the planarization layer was Hunt HPR-206 photoresist, which has a nominal thickness of 2 μm when spun at 5000 rpm for 30 sec. After patterning a TLR wafer with 1 μm steps in aluminum, the CD control over the steps was found to be rather poor. A second approach, which uses two coats of Shipley 1470 photoresist to form a 1.6 μm thick planarization layer, showed considerable improvement in planarization properties. This is demonstrated in Figures 5.8.1-5.8.2, showing good CD control in 0.5 μm lines over 1 μm steps in polysilicon, which

is much harder to planarize than aluminum steps, because of their nearly vertical edges. The latter process was therefore adopted to produce the planarization layer in this work. For geometry without severe topographical steps, one coat of 1470 (0.9 μm) can be used.

The postbake temperature of the planarization layer is critical. If the resist is baked at too low of a temperature, then outgassing in the planarization layer during the deposition of the isolation layer will occur, leading to defects in the pattern. A bake temperature of 225 °C was found to prevent outgassing. The wafers can not be used in a teflon cassette at this temperature, because the teflon outgasses and fluorine will be absorbed into the resist, disturbing the later etching process. Therefore, a quartz boat was used for the baking.

For the isolation layer, 500 Å of amorphous silicon was used. The machine currently used for silicon deposition is the MRC RIE-51. Two 150 mm wafers are placed on the smaller, bottom electrode and the TLR wafer is placed on the top electrode. An argon plasma sputters silicon from the bottom wafer, causing silicon to cover the TLR wafer. The standard practice for reactive ion etching is to place the wafer to be etched on the bottom electrode and introduce an etching gas into the chamber.

The imaging layer is 4,000 Å of Shipley AZ 2400-17 photoresist. The wafer is coated with a surface adhesion promoter, then 1.5 cc of filtered resist is dispensed on the wafer. After a prebake at 90 °C , the wafer is exposed on a Perkin-Elmer MEBES I with a dose of 20 $\mu C/cm^2$. The wafer is immersion developed and then postbaked at 120 °C .

After developing the imaging layer, the isolation layers and planarization layers are etched consecutively in the MRC-RIE system. The isolation layer is etched at 10 mTorr in a 20 watt SF_6 plasma for 3 minutes. The planarization layer is etched in a 5 mTorr, 200 watt oxygen plasma. Sufficient power is required during the etching to prevent redeposition of sputtered silicon from the isolation layer redepositing onto the planarization layer, inhibiting further etching of those areas. A small amount of fluorine can be bled into the system to prevent this, but at the expense of increased undercut. Thinner planarization layers (1 μm) were found to be free from these difficulties.

After the TLR has been used for patterning the substrate, it can be stripped with boiling sulfuric acid.

5.8.3 Contact Test Chip

The tri-level resist process was evaluated by fabricating many contact resistance structures. These test devices had a wide range of contact sizes, ranging from 0.25 × 0.25 μm^2 to 50 × 50 μm^2. This test pattern was chosen because contact resistance is believed to be a major limitation to submicron devices and the test structures are relatively easy to fabricate. By actually measuring submicron contacts, the device degradation resulting from contact resistance can be determined.

The major problem encountered was that of fabricating a wide variety of contact hole sizes, each with their correct size. The pattern is written on 0.25 μm addresses using a circular beam with the same diameter and a gaussian electron current distribution. Because of the gaussian profile, there is overlap between adjacent addresses, and hence the proximity effect [5.8.2]. Large contacts then require a smaller incident electron dose per pixel than small contacts to yield the same effective exposure.

Another problem results from shadowing during contact etching. During plasma etching, chemical radicals arriving from oblique angles are blocked by the sidewalls of the resist, and are prevented from reaching the small contact holes. As a result, small contacts are etched more slowly than large

ones.

The proximity and shadowing effects were compensated by using different doses for different contact hole sizes. A wafer with 1000 Å of LTO was coated with TLR, and then exposed with an array of different electron doses and contact hole sizes. After the TLR and LTO were etched and the resist was stripped, contact hole sizes were measured with an SEM and the optimal dose for each size was determined.

5.8.4 Experimental Procedure

A thin layer of pad oxide was grown on a lightly doped p-type Si ⟨100⟩ wafer. After a field threshold voltage adjustment implant and the pad oxide was stripped, a field oxide was grown (2,500 Å). A series of electron-beam global alignment marks were written, followed by a wet oxide etch and a silicon trench etch (3 μm) to delineate the marks. The global electron-beam alignment marks were written using a Perkin-Elmer MEBES I electron beam pattern generation system with a single layer of AZ 2400-17 resist.

The diffusion region mask and optical alignment marks were next transferred to the wafer using the MEBES I. A pad oxide was grown and the diffusion region was implanted with As (6×10^{15} at 100 kV). After an anneal to activate the implant, the pad oxide was stripped and a layer of 1,000 Å of LTO was deposited and densified. The contacts were exposed using the MEBES I with a tri-layer resist scheme described previously. The spot size was 0.25 μm. The contacts were etched using both dry etching followed by wet etching and totally dry etching with electrical end-point detection.

Contacts formed include pure Al, selective CVD W, and PtSi to arsenic doped junctions (R_s= 27Ω, X_j=0.35 μm, Ns≈ 2.5×10^{20}cm^{-3} ,. The metalizations of Al/Si was patterned using an Ultratech wafer stepper to give an overall mix-and-match lithography to enhance throughput. Contacts were also fabricated using optical exposure on all levels.

5.8.5 Results

The physical parameter which characterizes the interface between contact metal and semiconductor is the specific contact resistivity, ρ_c(Ω-cm^2). Values of $\rho_c$extracted from test structures of different sizes and designs often disagree by more than an order of magnitude, because simple 1-D models used for extraction of $\rho_c$can not account for parasitic fringing resistance associated with the 2-D current distribution around the periphery of the contact window [5.8.3]. In our previous work, 2-D modeling of contacts has shown good agreement with measured resistance for contact hole sizes down to 2 μm [5.8.4].

All three types of contacts exhibited deviations from 2-D models for contacts smaller than 1 μm (Figures 5.8.3-5.8.5). We believe this is due to the following: overetching of contacts required to ensure complete removal of SiO2, encroachment and worms in the case of CVD W (Figure 5.8.6), or consumption of silicon during silicidation. These can lead to a larger effective contact area which can be misinterpreted as a decrease in ρ_c. The overetching of the contact holes and/or silicidation increases the effective contact area (Figure 5.8.7). Because of the increased effective area of the contact, and the diffusion depth approaching the size of the contact hole, the assumptions implicit in the 2-D modeling become invalid. Further modeling of the current flow will require a 3-D simulation.

The value of ρ_c was extracted from the contact resistance data using the 2-D model in the regime where the 2-D assumptions are still valid. In the case of Al, we have found that by optimizing the annealing conditions, if Si dissolution in Al and Si precipitation at the Al/Si interface can be avoided, extremely low values of ρ_c between 1.45×10^{-8} and $1.6 \times 10^{-10}\Omega$-cm^2 can be obtained. These are at least a factor of 50 lower than previously reported [5.8.5]. For $0.5 \times 0.5\mu$m^2 contacts of Al, W, and PtSi the contact resistances measured by the cross bridge kelvin resistor technique were 11, 20, and 31 Ω, respectively. For $0.25 \times 0.25\mu$m^2 contacts, the contact resistance for PtSi was 44 Ω. These values are the lowest ever reported.

5.8.6 Future Work

We are in the process of fabricating additional contact resistance wafers including those with boron junctions. We are also fabricating a number of pMOS devices using a hybrid lithography scheme with the MEBES I used to define the actual gate width, and the Ultratech stepper to expose large areas of the wafer and hence increase throughput.

5.8.7 References

[5.8.1] Brodie and Muray, *The Physics of Microfabrication*, New York: Plenum Press, 1984, p. 307.

[5.8.2] Sze, *VLSI Technology*, New York: McGraw-Hill, 1983, p. 286.

[5.8.3] W. Loh, K. Saraswat and R. Dutton, *IEEE Elect. Dev. Lett.*, vol EDl-6, p.105,1985.

[5.8.4] W. Loh et al., *IEDM Tech. Dig.*, 1985, p.586.

[5.8.5] R. L. Maddox, *IEEE Trans. Elect. Dev.*, vol ED-32, No. 3, p. 683, 1985.

5.8.8 Figures

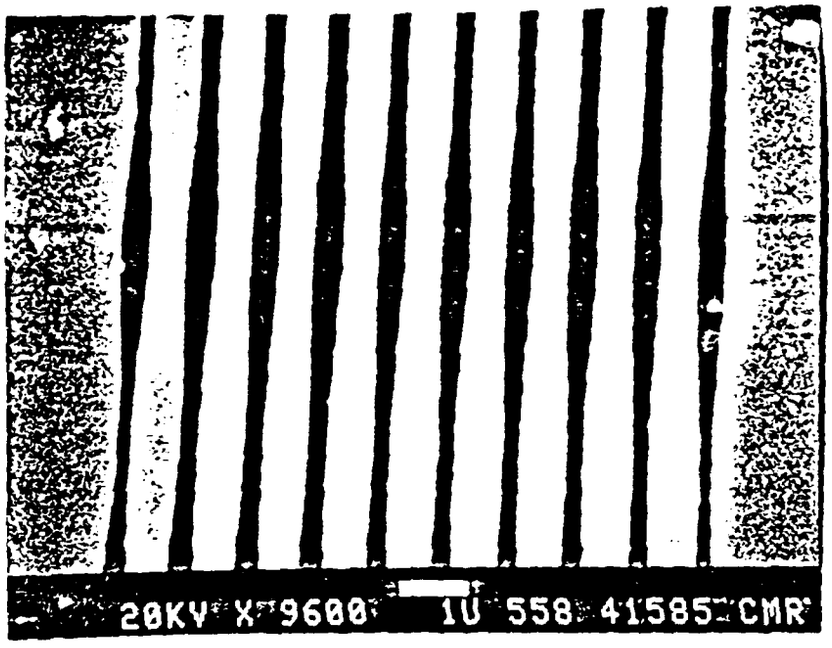


Figure 5.8.1: 0.5 μm lines in TLR over 1 μm trenches in aluminum

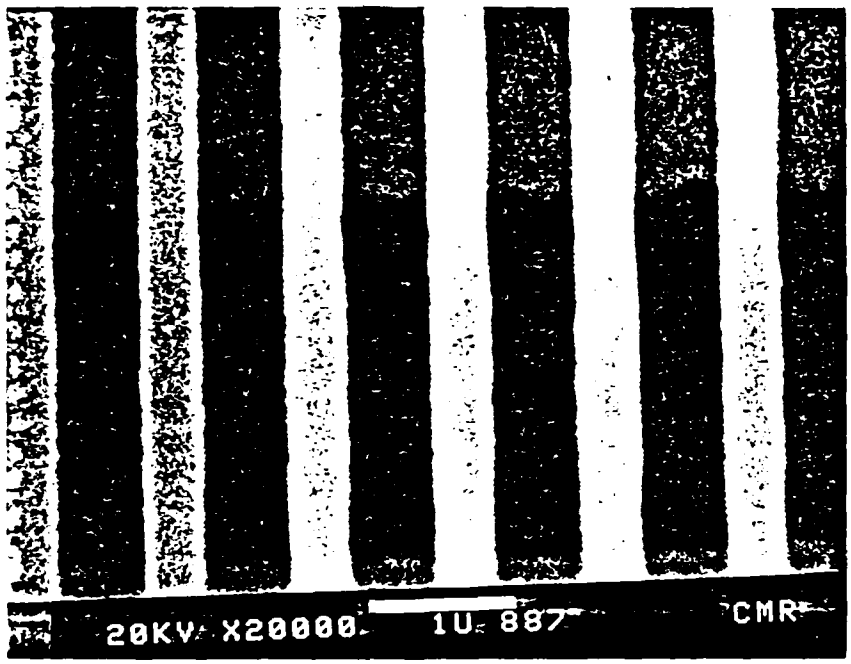


Figure 5.8.2: 0.5 μm lines in TLR over 1 μm trenches in polysilicon

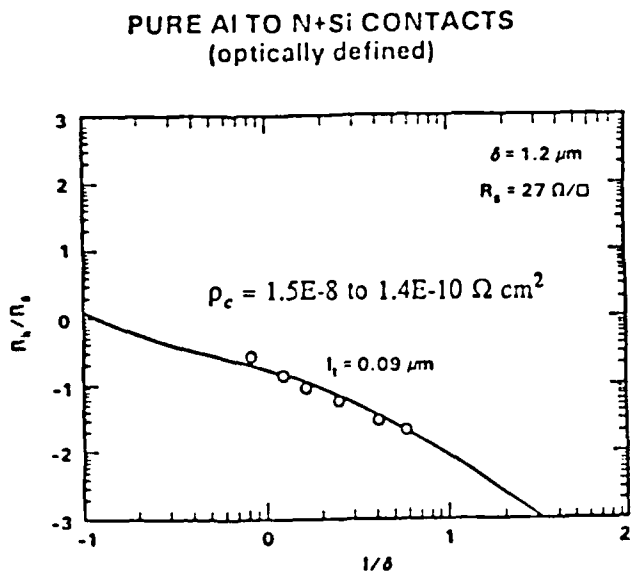


Figure 5.8.3: Normalized Contact Resistance for Pure Al to n+ Si

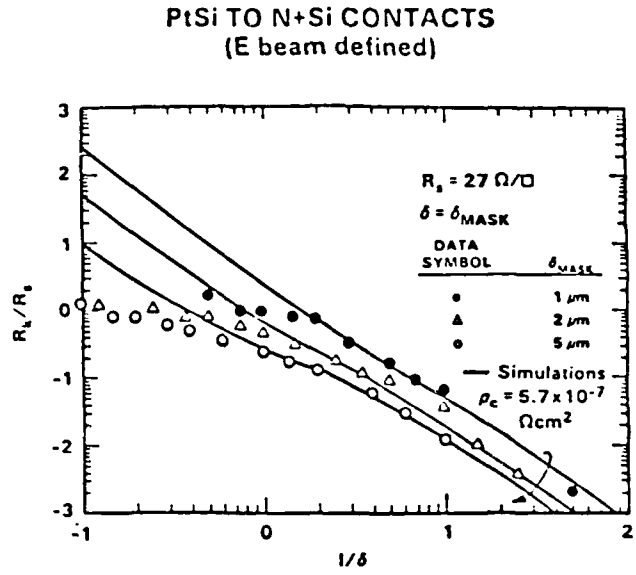


Figure 5.8.4: Normalized Contact Resistance for PtSi to n+ Si

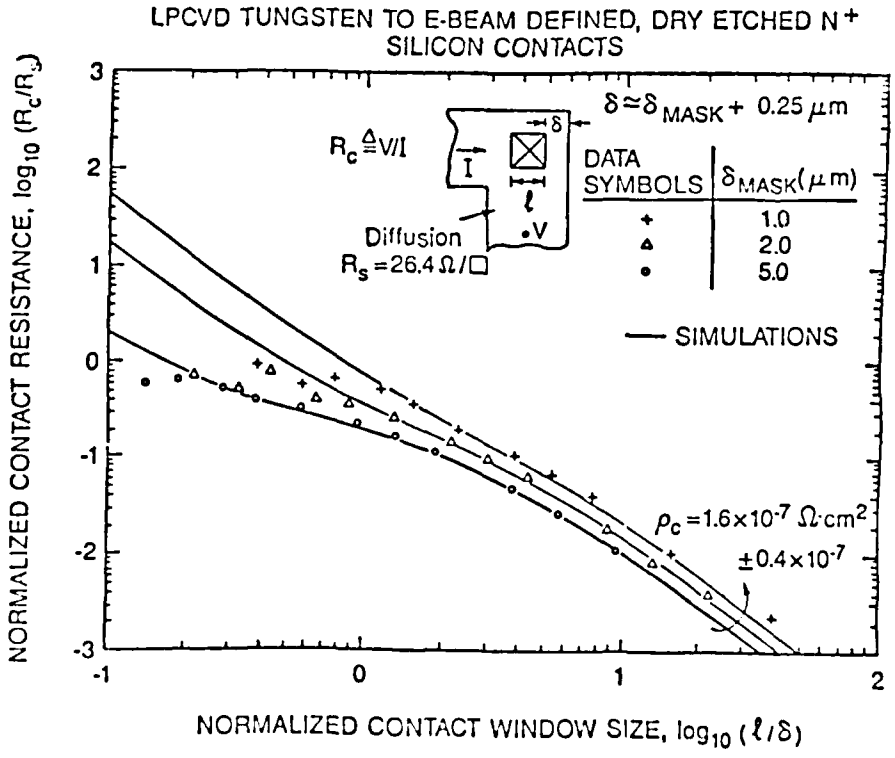


Figure 5.8.5: Normalized Contact Resistance for CVD W to n+ Si

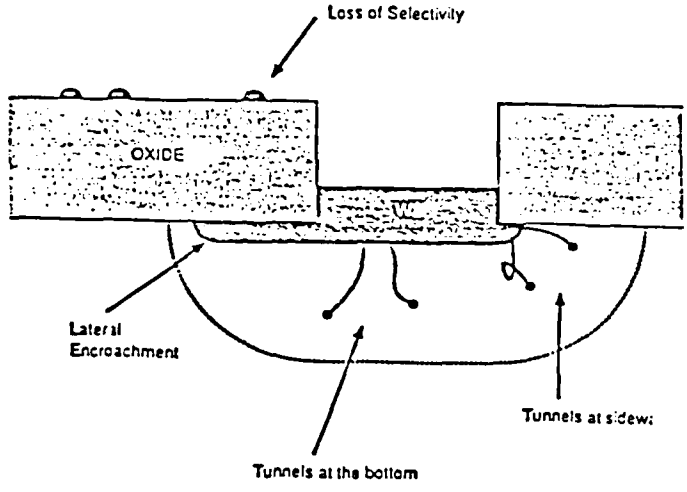


Figure 5.8.6: Problems with Selective CVD W

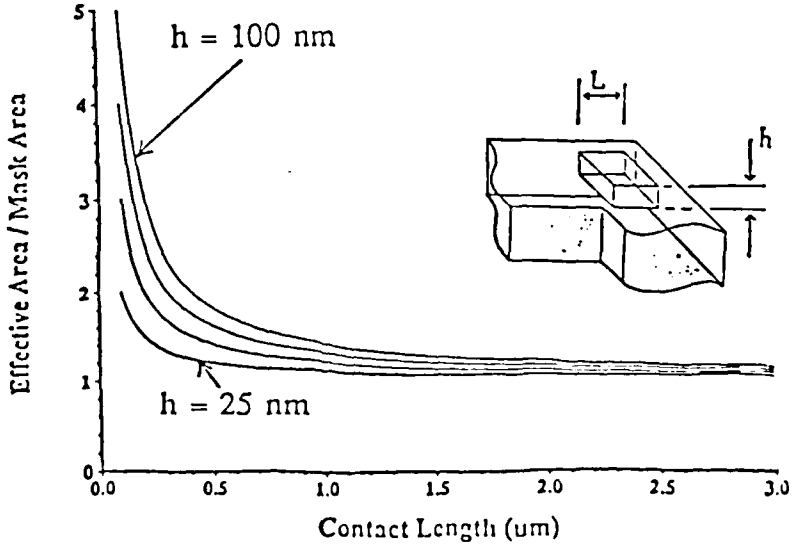


Figure 5.8.7: Relative Increase in Effective Contact Area with Overetching

5.9 Ion Implantation Modeling

Lawrence Rauchwerger and Krishna Saraswat

5.9.1 Introduction

The work on the ion implanter is part of the more general effort to model semiconductor equipment for the fab modeling project. Equipment models are useful for several purposes:

1. As a module for an off-line simulator of the fab.; It would be able to predict yields, throughput and other general production variables.

2. As an off-line prediction program for device and process parameters for that particular process step.

3. As a design tool for semiconductor equipment.

4. As an on-line control program for the equipment itself as well as for the process.

Constructing a model for an ion implanter means finding a relation between the different input process variables and the resulting implant characteristics of the wafer. It requires an almost complete knowledge of the interdependences between the different process and equipment variables. A simulation of the ion implanter may be obtained only after a lot of analytical modeling and experimental verification.

5.9.2 Experimental Work

This past year we have received as a donation from Varian Associates Inc. a 350D ion implanter that has been installed and is now fully operational. We have performed characterization experiments and compared the results with those obtained on other, similar, implanters within the University and local industry. Due to the requirements of our integrated circuits laboratory we have had converted the implanter endstation to a 3 inch one. Unfortunately we did not have the corner faraday cups changed accordingly and the uniformity monitor reading can not be useful. With the move of the IC Lab into its new facility in the Center for Integrated Systems we will upgrade the whole process to 4 inches, so this problem will be taken care of. We have been trained to operate and service the ion implanter. The characterization experiments which were performed reguarded mostly dose accuracy; It was very important to establish a correspondence between our old ion implanter and the new 350D.

Ion Implanter as a Research Tool

As part of an effort to improve the metalization process in our fab we have studied the glass reflow and step coverage.

Purpose of the Experiment Phosphosilicate glass (PSG) is widely used in the fabrication of the integrated circuits to provide passivisation and electrical insulation between metal interconnects and underlying structures. PSG deposition tends to leave sharp profiles which are difficult to cover uniformly with subsequent metal layers. In addition, etched vias made to contact underlying layers can also exhibit profiles that are difficult to cover uniformly. Thus a smoothing of the PSG surface topology is often needed before metal deposition. In conventional MOS processing, oxide flow (or reflow) has been extensively used for improving metal step coverage over PSG. This is achieved in a furnace by annealing the PSG film containing 6 to 8at 1000°C - 1100°C for 20-30 minutes in an N2 ambient. These long, high-temperature cycles make it difficult to maintain shallow junction depths in VLSI devices. With current need for fine geometries and shallower junctions, techniques to reduce the temperature and/or length of the reflow cycle become essential. J. McVittie and others have tried successfully to reduce the reflow temperature by using a steam ambient. By using rapid thermal annealing in steam they have obtained a lowering of the reflow temperature by about 50°C .

We have proposed to lower even more this reflow temperature by ion implanting the glass before the anneal cycle.

Experimental Procedure The following test structure has been processed:

1. Grow 1000 Å oxide on a Si wafer
2. Deposit 5000 Å polysilicon
3. Dope poly with POCl5
4. Pattern poly with a test mask
5. LTO (doped with 7% P) deposition (5500 Å)
6. Implant As under variable conditions
7. Reflow under variable conditions
8. Prepare samples for the SEM (cut the samples and etch in SF6 for 45 sec)
9. Take SEM pictures, measure the reflow angle and tabulate the data.

Step 6. and 7. will be iterated for different conditions of implant and reflow in order to generate data for a factorial analysis.

The reflow has been done with a Heatpulse rapid thermal annealer, in the case of the RTA and in a furnace of our lab for the classical anneal cycle.

Also important to mention is that we defined as an acceptable reflow result when the angle formed by the horizontal with the tangent to the glass covered sidewall of the polystep is less than 75°.

Results and Discussion From all the possibilities we have chosen to implant As because different articles in the literature mentioning it, as being the most effective specie.

Implant Energy First we have tried more implant energies, keeping dose (3E16) and RTA conditions(1050°C , 40 seconds, N2 ambient) constant - 60 KeV, 80 KeV, 120 KeV, 150 KeV and 200 KeV. The best reflow has been obtained for the lower energies but from range versus lost depth during cleaning procedure considerations, we have decided to continue the experiments at 80 KeV. The best result under these conditions was a reflow angle of 87 degrees. A triple implant i.e. 40 KeV, 120 KeV, 200 KeV has been tried too, but no better results have obtained.

Implant Dose Dose variations between 1E16, 3E16 and a triple implant i.e. 3E16 at the three before mentioned energies have been tried. At 1E16 almost no reflow has been obtained. 3 X 3E16 was the best (65 degrees) but besides highlighting a trend, it is a rather impractical dose (long implant time, high As concentration). We have chosen 3E16 as reference dose.

RAPID THERMAL ANNEALING

Temperature, Time We have varied the temperature between 1050°C and 1100°C and the time between 50 seconds and 90 seconds. Not surprisingly, we could conclude that the higher the temperature, the better the reflow. We obtained for the standard dose and energy an angle of 85 degrees at 1050°C /50sec and an angle of 70°at 1100°C /90sec. The last result was the best ever to be obtained during the whole experimental work.

Environment (Gas) Four types of annealing environments have been experimented: N,O, Forming gas, Ar. No real difference, between these four conditions has been observed.

Discussion To actually measure the effectiveness of improving anneal conditions by ion implanting we compared our results with a control wafer that has been subjected to the same anneal conditions but no implant. The result was disappointing: One can improve (lower) the reflow angle by ion implantation by a maximum of 10 degrees, which does not translate into a significant lowering of the RTA temperature and/or time.

An increase in P concentration for the glass will have a much more significant effect than ion implanting; By using an 8% PSG one can obtain very acceptable results for the presently used geometries. However a high percentage PSG is generally not welcome (it may increase diffusivity of other layers and increase the junction movement, offsetting the advantage of better anneal conditions; phosphorus may also diffuse itself into the lower layers)

FURNACE ANNEALING We have also tried to better the furnace anneal conditions by ion implanting As into the PSG . The different parameters have been varied in a very similar manner as in the RTA case (implant conditions were kept identical)

Higher implant energies seem to favorize a better reflow.

At 950°C /5 minutes in steam ambient excellent results have been obtained. A summary of all the obtained data is reproduced in table I. (To the 5 minutes in furnace we must add about 20-25 minutes of ramping)

Discussion Making the same comparison with a control wafer as in the RTA case we could report a significant difference between implanted and not implanted glass. The reflow angle is lowered by as much 22°which translates in temperature differences of maybe 50°C in anneal temperature. Therefore in the furnace anneal case it makes sense to introduce an extra implant cycle in the processing of VLSI.

PSG / time [min]	8%	7%	7% + implanted
5	52	87	70
10	40	77	45
20	30	68	34
40	19	47	25

Implant with As: Q=3E16/E=150keV

Figure 5.9.1: Reflow angle for anneal in furnace at 950°C

Conclusion The difference between the two anneal methods with respect to ion implanting could have the following explanation: As implantation brakes bonds in the already amorphous glass lowering the cohesion forces, maybe introducing some extra mechanical stress and in general reducing the viscosity of the PSG. Also a higher percentage of P will probably make the glass a less stable compound as opposed to purely homogeneous $SiO-2$. But these considerations are valid for both RTA and furnace anneal cycles. The fact that not higher temperatures but longer times make reflow more complete for the same type of PSG leads to the idea that, we have a mass transport problem. It seems that the temperature beyond a certain value does not increase the glass velocity in the direction of flow in a significant manner; temperature facilitates, but is not the driving force behind the flow (surface tension probably is). Therefore for the same flow velocity, time will be the main factor for the spatial redistribution of the glass mass. In the case of the RTA, ion implantation of As will actually have the same effect as in the furnace anneal case but, because the time is so short, no improvement versus the nonimplanted case can be measured. Also, at high temperatures, the number of bonds broken by the ion implantation is small compared to htose broken by temperature.

Ion implantation is useless and would probably be as well for BPSG, although it has the intrinsic advantage of a lower reflow temperature. However, ion implantation makes sense for furnace annealing.

We conclude that the use of RTA is restricted to VLSI applications where high phosphorus content is tolerable in PSG films. Also the use of steam environment in RTA,as shown by McVittie, is a condition to make reflow temperature acceptable; unfortunately it is not practical.

In general, we would like to make the remark that RTP has been developed for use in processes where dopant or in general, particles, redistribution should be minimized.

5.9.3 Future Work

As mentioned in the introduction modeling of the ion implanter is our ultimate goal. And we would start by defining a set of input and output variables that would frame the model both for implanter studies as well as for the fab simulation.

Input variables We think that input variables to the implanter seen as a "black box" are the known process variables: (Input Process):

- IP.1. dose :nominal value and tolerance
- IP.2. energy: nominal value and tolerance
- IP.3. ion specie because of the thermal and charging effects in high current implanters we should add to this input list
- IP.4. maximum allowed temperature X time product (maximum allowed dopant redistribution for the implant cycle and maximum self annealing degree)
- IP.5. maximum permanent charging degree
- IP.6. also for the fab tracking system we would need a cassette ID (which could point to an entry in a data base)

The precise meaning of the variable IP.5 has yet to be determined.

These process defined variables are joined at the implanter model input by equipment parameters. A partial list of these specific inputs could be the following: (Input Implantation)

:

- II.1. Source control settings
  - Source gas pressure
  - Arc current
  - Filament current
  - Temperature (for vaporizer)
  - Beam steering control
  - Source magnet current
  - Analyzer magnet current
  - Beam Filter on/off status
  - Terminal elapsed time

There might be other parameters that could indicate lifetime of the source and history of what source gases have been used.

- II.2. Beam scanning parameters (lenses, acceleration voltage, XY scan, etc)
- II.3. Beam conditioning (Electron Flood parameters)
- II.4. Endstation settings (Target pressure, dose measurement mode, etc)
- II.5. VLSI endstation wafer handling settings (cassette ID, etc)
- II.6. Monitor and Control Mode of the Implanter (Dose, Uniformity meters, software control mode of the main computer, etc)
- II.7. Interlock system
- II.8. Auxiliary variables (power, compressed air, liquid N2, exhaust- ventilation)

The tolerances of all mentioned inputs are to specified too.

The history of significant parameters in the form of trends can be a very useful information; Some modern implanters generate this information with their own software.

Output Variables Output variables can again be defined at the equipment level and, on end result, on the wafer level.

Some Implanter outputs would be: (Output Implanter):

- OI.1. Beam Current characteristics:
  - Average value (in time)
  - Peak value (in time)
  - Spatial distribution
  - Degree of neutrality (obtained or measured)
  - Degree of purity (species, energy)
- OI.2. Wafer Temperature (measured) spatial and temporal distributions
- OI.3. Wafer Charging Curve This output has yet to be properly understood.
- OI.4. Source Temporal Variations(during implant)
- OI.5. Vacuum Temporal Variations(during implant)
- OI.6. Interlock Status (during implant)
- OI.7. Monitor and Control Results (Dosimetry, Uniformity Meter)

Process Outputs deal mostly with the end result of the implant cycle as measured on the wafer. Examples are: (Output Process):

For each device on the wafer we can distinguish:

- OP.1. Implant Profile Characteristics (Range, Straggle, Channeling Tail)

- OP.2. Damage Profile Characteristics (Amorphisation, Self annealing)
- OP.3. Charging Characteristics
- OP.4. Impurity Implant (species, quantity due to imperfect implant conditions)

For all the above local parameters a

- OP.5. Uniformity of the Local Parameters (spatial)

can be defined.

Conclusion The list of the so structured input and outputs has to be completed and brought to a common denominator with that one established by the "FABLE" research group. We have made contacts in that direction with Byron Davies, who heads the FABLE project.

Another effort will go into a better instrumentation of the implanter. And finally the equipment "transfer functions" will be established. This will enable us to simulate and predict any process cycle on this piece of equipment.

5.9.4 References

[5.9.1] J.F.Ziegler, "Ion Implantation", Academic Press, 1984

[5.9.2] "Varian 350D Ion Implanter Manual", 1985

[5.9.3] J.F. Gibbons, "Ion Implantation in Semiconductors", 1972

[5.9.4] Heiner Ryssel, "Ion Implantation for Very Large Scale Integration", Advances in Electronics and Electron Physics, vol. 58, Academic Press, 1982

[5.9.5] Devereaux, Chen, Szeto, Hornq-Sen Fu, "Enhanced Flow of Phosphosilicate glass by Ion Implantation", IEEE, 1983

[5.9.6] Mercier, Beerkens, Calder, Naguib, "Rapid Thermal Reflow Of PSG Films", ECM

[5.9.7] Blech, "Step Coverage by Vapor Deposited Thin Aluminum Films", Solid State Technology, December 1983

[5.9.8] Box, Hunter, Hunter, "Statistics for Experimenters", Wiley, 1978

Chapter 6

Diagnostics and Yield Modeling

During the 1986 fiscal year, the Diagnostics and Yield Modeling task concentrated on the development of in-process monitors for photolithography and ion implantation; the development of a comprehensive end-of-process test vehicle for CMOS process problem debugging and yield prediction experiments; establishing a first-rate electrical characterization facility; and the development of test software required to support the testing and characterization activities. In the course of these efforts, we have secured a donation from Hewlett-Packard of a complete HP4062A-based electrical characterization system, and Rucker & Kolls, Micromanipulator, and Suss probers, and have established a collaborative testing effort with Hewlett Packard Laboratories. Dr. Dirk Bartelink is the industrial mentor of this activity. Moreover, we have secured donations of the ENHANSYS and RS/1 statistical analysis packages to support the data analysis tasks associated with the above activities.

The research in in-process monitors has led to the development of an electrically testable photolithography monitor [8.1.4] capable of separating the photolithographic contribution to patterning defects from the pattern transfer (etching) contribution by employing the photoresist developer as an electrolytic etchant for the underlying metal film test structures. The technique has been successfully applied to the evaluation of pattern recovery in the triplicated exposure scheme [8.1.2].

The work in ion-implant monitors, which resulted in a high spatial resolution monitor for dose uniformity [8.1.1], has been expanded to include monitors for channeling and shadowing, as well. A process flow has been developed and mask set generated to integrate the dose, channeling, and shadowing monitors on a single wafer to provide simultaneous evaluation of implanter performance. This work is described below by Anthony McCarthy.

The work in IC end-of-process monitors, on the other hand, employed a systematic procedure for decomposing the complex IC structures into sets of simple, unambiguously interpretable structures to develop a comprehensive, logical approach to process problem debugging via determination of the structural location and magnitudes of electrical faults [8.2.1]. This work is described below by Willie Yarbrough. (We have conducted 12 invited seminars on this topic in industrial facilities and conferences throughout the U. S., and a complete set of end-of-process monitors modeled after this work has been implemented in at least one industrial facility in the U. S.)

To complement the above effort, we have also initiated a study to determine the test structures required to empirically decompose the parameters employed in circuit simulation (the "process file") into their process variables. Like Yarbrough's work, which separates the composite defect density into its numerous structural components, this work aims to decompose the SPICE parameter distributions into their process variable distributions to identify and rank the process steps most

responsible for IC performance variability and, hence, in need of improved control. A preliminary study of an AI assisted approach to the analysis of the parametric data that would be collected for this purpose has been carried out as a course project and is described below by Tsu-Chang Lee.

To complement the work in diagnostic techniques, we have also initiated an effort to develop a set of alignment, CD, and defect density assessment test structures for the purpose of implementing a statistical, empirically based approach to the development of circuit layout guidelines. Along with yield prediction models, these would serve in place of the inflexible "design rules", allowing a circuit designer to assess the impact on circuit yield of "relaxed" vs "aggressive" layout rules, thereby permitting product performance/cost optimization. The initial efforts in this domain are described below by Greg Freeman.

The work of William Loh presents a unified approach for the accurate extraction of specific contact resistivity for ohmic contacts from measured contact resistance using the Cross Bridge Kelvin Resistor, the Contact End Resistor, and the Transmission Line Tap Resistor. Conventional 1-D models overestimate the value of specific contact resistivity because of a parasitic resistance due to two-dimensional current crowding around the periphery of the contact. Using 2-D simulations we have accurately modeled the current crowding effects and have extracted values of specific contact resistivity independently of contact size and the test structure type. For each particular structure, a universal set of curves has been derived, which, given the geometry of the structure, allows accurate determination of , without the actual use of the 2-D simulator. Accurate values of specific contact resistivity for various contact materials to heavily doped Si have been determined. The data confirms that in the past researchers have overestimated specific contact resistivity and that it will not limit device performance even with submicron design rules.

6.1 In-Process Testing

Anthony M. McCarthy and Wes Lukaszek

6.1.1 Ion implant monitor dosimeter

A new Ion Implant Monitor test structure has been proposed. A direct measurement of the sheet resistance of the implanted species is employed. This has the advantage of unambiguous verification of the electrical performance of the layer. The proposed method has a high spatial resolution, approximately 300μm between measurement locations. The technique does not suffer from the problem of compensation for non-infinite planes associated with wafer edges in four point probe measurements. Voltage measurements are directly converted to sheet resistance, thus measurements may be performed rapidly. Experiments were performed for Boron only since this element is most commonly employed in CMOS threshold adjust implants. The test structure was verified to operate in the range $7\times10^{11}/cm^2$ - $1\times10^{15}/cm^2$. This work is described in section 6.1.4.

6.1.2 A novel technique for consecutive pattern developing and etching of thin aluminum films

A novel technique has been developed to enable consecutive developing and etching of thin aluminum layers. In VLSI patterning, after the photoresist has been delineated by the developer the etch rate of the underlying aluminum due to the developer is negligible (approximately 100 Angstroms/minute). The proposed method alters the mode of operation of the developer to electrolytic action by bringing an electrode in contact with the developer once the develop cycle is complete. This accelerates the etch rate of the underlying aluminum to as much as 2500 Angstroms/minute. This etching technique has applications in VLSI technology for rapid generation of electrical test structures, detecting particles in resist, and monitoring linewidth and alignment.

It was the purpose of this research to explore the possibility of determining the contribution of defects due to photolithographic processes alone. Previous efforts at determining this employed a two step process where the pattern was first exposed on the wafer and then in a "developer" solution, followed by pattern etching. The normal means of detecting defects in pattern transfer techniques is to define electrical test structures consisting of serpentines and interdigitated combs. Electrical measurement of these structures permit detection of shorts in the combs and opens in the serpentines. This research introduced a method whereby developing and etching are performed as consecutive steps in the same solution. It is claimed that this method enables the detection of the true contribution to the defect count of lithographic steps alone.

The developer used in these experiments, Shipley MF312, is one in common use for VLSI processing. The photoresist used was Shipley AZ1470.

This technique was applied to a voting lithography scheme. Voting lithography is a procedure for elimination of mask induced defects without resorting to expensive mask correcting equipment.[6.1.2] The etching technique was employed to determine the defect densities associated with various means of employing the method and eliminating the contribution due to etching in a second solution which would introduce an unwanted, but otherwise unavoidable, variable into the experiment. A detailed examination of the implications of voting lithography for a VLSI patterning technology was given in a recent publication.[6.1.3] Detailed conclusions extracted from the etching experiments, not reported in the above paper, are given in section 6.1.5.

6.1.3 Ion implant monitors for dosimetry, channeling, shadowing.

The effect of wafer orientation, angle and tilt on ion channeling in Silicon has been empirically observed to be minimized for particular values of these parameters.[6.1.5] Shadowing is the result of the inadvertent blocking of an implant due to the existence of "tall" features on the wafer. A set of ion implant monitor electrical test structures designed to measure dose uniformity, channeling and shadowing effects for the purpose of implanter calibration and evaluation will be integrated onto one wafer. This will enable the simultaneous monitoring of these effects in a single implant.

The mask set has been designed as a generic set such that the same mask set and similar process may be used in performing experiments for p-type and n-type implants, although the substrate and epitaxial layer polarities need to be reversed. No metallization is required. Probing is performed by using standard probes on heavily doped contact regions.

The substrate used in the initial experiments is p-type with a n-type epitaxial layer. Isolation for the structures is achieved by forming mesas using KOH. For illustration purposes in the

accompanying figures a light dose p-type implant is used.

The dosimetry structure uses a Van der Pauw configuration in which the central portion contains the light dose implant and the contact pad areas are heavily implanted, as shown in fig. 1. This is similar in principle to the structure outlined in [6.1.1].

The shadowing structure uses a self-aligned Tee composed of oxide, nitride and poly, as shown in fig. 2. The oxide is used to block the implant and provide the shadow. Electrical linewidth measurements are performed to determine the extent of assymetry in the implanted regions. Sheet resistance information for the layer will already have been obtained using the previous structure.

The channeling structure uses the JFET technique with the implanted dose acting as the control gate, as shown in fig. 3. Three different structures are implemented in this process. Structure (a) provides information on the junction depth when the implant impinges directly on the silicon surface. Structure (b) randomizes the incoming implant with a dielectric screen. Structure (c) is the reference structure from which the local epitaxial layer uniformity may be extracted. The monitored implanted is blocked in this structure.

6.1.4 A new Ion Implant Monitor Electrical Test Structure

Introduction

In this paper, a new Ion Implant Monitor test structure and measurement method is reported. A direct measurement of the sheet resistance of the implanted species is employed. This monitor has the advantage of unambiguous verification of the electrical performance of the layer. The proposed method has a high spatial resolution, approximately 300μm between measurement locations and has been shown to have resolution down to at least $7\times10^{11}/cm^2$. The technique does not suffer from the problem of compensation for non-infinite planes associated with wafer edges as in four point probe measurements. Voltage measurements are directly converted to sheet resistance, thus measurements may be performed rapidly.

Structure Description and preparation

The structure is implemented on a wafer of opposite polarity to the type of implant being measured. The plan and profile view is shown in fig. 1. Only Boron implant measurements are reported in this paper. Three inch wafers of a high resistivity phosphorous doped Silicon, in the range 10 - 20 Ω-cm, were selected. A thermal oxide layer of 1000 Angstroms thickness was grown. 5000 Angstroms of undoped Polycrystalline Silicon was deposited to act as a blocking layer. Mask number one was used, fig. 4, to define probe pad areas by etching openings in the Poly. A moderate energy heavy dose of Boron of $1\times10^{15}/cm^2$ was implanted through the oxide. Mask number two was used to define the cross-bridge areas for the second implant, fig. 5. After a second Polysilicon etch and resist strip the wafers were ready for implantation and stored until time of use. The procedure following the second implant was standard: the wafers were annealed in an Argon ambient at 900C for thirty minutes, and the oxide was stripped. Measurements were performed using a standard parametric system and automatic wafer probe with no special probes or software.

Results

Wafers were implanted with doses in the range $7\times10^{11}/cm^2$ - $1\times10^{15}/cm^2$. Wafer maps of the results taken from wafers implanted with doses of $1\times10^{13}/cm^2$, $1\times10^{12}/cm^2$ and $7\times10^{11}/cm^2$ are

shown in fig. 6, 7 and and 8 respectively. The standard deviations are 2.2, 2.2 and 2.7% respectively, which is very good for low doses. A plot of measured sheet resistance versus dose, shown in fig. 9, agrees well with previously published data.

For each dose one wafer was selected to measure data within a square die area of approximately .5cm on a side. On each wafer twenty eight sites on each of three die were measured. The die lay on a diagonal of about 60mm in length on the three inch wafer. In the case of the $1\times10^{13}/cm^2$ dose, the die area was approximately .25cm on a side, the number of test sites was twenty per die and the diagonal length was about 50mm. A summary of the within-chip measurements is given in Table 1. The standard deviations are very tight as expected. The initial heavy implant was also monitored. It can be seen from the table that it was multi-modal in the range 110-140 ohms/square and this does not affect the low dose measurements.

Conclusions

A new electrical ion implant monitor has been presented, and tested for measuring implant doses as low as $7\times10^{11}/cm^2$. The technique employs the Van der Pauw structure to measure the sheet resistance of implanted layers. It is a direct confirmation of the electrical performance of the implant. The spatial resolution is an order of magnitude greater than currently available implant monitoring systems. This permits the possibility of measuring statistics within a die area. The measurement does not require compensation for non-infinite planes at wafer edges, and no special hardware or software is required other than a standard automatic parametric measurement system.

6.1.5 A novel technique for consecutive pattern developing and etching of thin aluminum films

Abstract

A novel technique has been developed to enable consecutive developing and etching of thin aluminum layers. In VLSI patterning, after photoresist has been delineated by developer, the etch rate of underlying aluminum by the developer is approximately 100 Angstroms/minute. The proposed method alters the mode of operation of the developer to electrolytic action by bringing an electrode in contact with the developer once the develop cycle is complete. This accelerates the etch rate of the underlying aluminum to as much as 2500 Angstroms/minute. The technique has been applied to the evaluation of defect densities associated strictly with the photolithographic process in a voting lithography scheme. The technique unambiguously verifies the benefits of voting lithography in eliminating random defects in VLSI masks.

Introduction

It is a consistent goal of IC fabrication facilities to minimize the introduction of defects onto wafers during processing. The purpose of this paper is to explore the contribution of defects due to photolithographic processes alone. Previous efforts at determining this contribution in electrical test structures employed a two step process where the exposed pattern was developed in a "developer" solution followed by pattern transfer via etching.

The most common means of detecting defects in pattern transfer techniques is to define electrical test structures consisting of serpentines and interdigitated combs. Electrical measurements on these structures permit detection of shorts in the combs and opens in the serpentines. Because the

measurements are done after the completion of the two-step process, this renders the separation of the individual contributions due to lithography and etching impossible. This paper introduces a method whereby developing and etching are performed consecutively in the same solution. It is claimed that this method enables the detection of defects introduced by the lithographic steps alone.

The idea proposed in this paper is to employ the developer to both develop the exposed pattern and etch the underlying metal layer. The prepared wafer is immersed in developer solution. After completion of the develop cycle electrolyte power is provided by applying potential to the wafer back-plane and to an electrode in the solution. As current passes through the developer and the wafer through the open areas in the resist, the metal is etched and the patterns are delineated in the metal layer. The developer used in these experiments, Shipley MF312, is one in common use for VLSI processing. The metal selected for the underlying patterns was aluminum. The photoresist used was Shipley AZ1470. At room temperature the aluminum etch rate in Shipley MF312 is approximately 100 Angstroms per minute. The dissolution of unexposed photoresist by developer is greater than 1000 Angstroms per minute. By electrolytically accelerating the aluminum etch rate to 1000 Angstroms per minute, it is possible to transfer the pattern with the photoresist.

Sample preparation

Though the procedure described above is a general one, the particular application proposed in this paper is the detection of defects in photoresist introduced by the photolithographic process. For this purpose special attention must be given the material structure of the wafer sample. It must permit vertical passage of current through the wafer during the etching process and also permit electrical isolation between test structures during testing. This latter requirement is particularly severe in the case of interdigitated comb structures where the long lengths of closely spaced lines demand very low lateral leakage currents.

This problem was solved by first depositing a thin layer of 2000 Angstroms of LPCVD undoped polycrystalline silicon on the wafer followed by 1000 Angstroms of e-beam evaporated aluminum. The substrates used were p-type silicon 0.1 - 0.5 Ω-cm. This structure permits current conduction through the wafer but very little current conduction across the wafer between adjacent metal patterns defined on the polysilicon. Backside polysilicon was removed prior to the deposition of 1μm of aluminum on the back of the wafer.

A cross-section of the sample is shown in fig. 10. The thin film processing was performed as follows.

1. Wafer: p-type, $\langle 100 \rangle$, approximately .1 Ω-cm.

2. Polysilicon deposition: LPCVD, 600C, 2000 Angstroms.

3. Backside polysilicon removal.

4. Backside e-beam aluminum evaporation, 1μm.

5. Frontside e-beam aluminum evaporation, 1000 Angstroms.

6. Photoresist spin, Shipley AZ1470, prebake 95C, 25 minutes.

7. Pattern exposure: Ultratech model 900, energy 110mJ/cm^2

Experimental setup

A beaker of developer was used in which a platinum electrode and the chuck holding the wafer were immersed, as shown in fig. 11. The wafer was held in position by vacuum and contact was made to the backside of the wafer through the metal chuck. Negative potential was applied to the electrode and positive potential to the chuck. Contact to the chuck was made by a hollow metal tube. The metal tube was isolated from the solution by plastic insulation. The teflon material surrounding the chuck served to insulate the chuck from the solution. A resistor was placed in series with the power supply to provide for a voltage tap when using a chart recorder to monitor process end-point.

After the photoresist is developed the power supply provides a constant current flow of 200ma-450ma through a three inch wafer. Bias is reduced to just maintain constant current initially. Once the metal patterns begin to be delineated the current decreases as the power supply voltage remains constant.

Because the current flowing through the wafer at constant voltage decreases once the exposed aluminum is etched away, end-point is readily monitored. An example of the I-V curve for end-point detection is shown in fig. 12.

Electrical Measurements

Using patterns defined in aluminum, as shown in fig. 13. it is possible to test for particles in the resist by electrical tests for short and open circuits. If a particle in the resist bridges a space, the aluminum will not be etched out under it resulting in a short circuit between two lines of the pattern. Similarly, if a void in the resist exists over a line, the metal line is etched out resulting in an open circuit. These patterns may be delineated at various dimensions of line and space to determine defect densities at the selected dimensions.

Advantages over previous methods

The major advantage of this method of metal etching is that the developing of the photoresist and etching of the metal are performed in the same solution. This enables a more reliable detection of defects due to photoresist processing alone.

A second significant advantage is the availability of end-point detection. There may be as much as a factor of ten reduction in current level thus enabling easy end-point detection and control. This indicates an ease of reproducibility which is not readily available with other etching techniques. New features of this method of etching include:

1. The etching of metal using resist developer as an electrolyte.

2. The decoupling of particles introduced by metal etching by performing the etch in resist developer.

3. The novel use of polysilicon as a conductor in the vertical dimension and as an insulator in the horizontal direction.

4. End-point detection is possible because current is passed through the wafer.

Application of the technique to voting lithography

In the voting lithography scheme, a number of mask fields containing nominally identical patterns are aligned and exposed in sequence at the same site, each with an equal fraction of the nominal exposure dose. The resulting optical intensity that impinges on the resist is the sum of these exposures. Consequently, in a scheme using Nfields, a random mask defect unique to a single participating field, i.e. a nonrepeating defect, affects only 1/N of the total exposure. The effect of this exposure deviation can be minimized with a properly adjusted exposure dose and an adequate resist contrast. It is therefore possible to produce essentially defect-free resist patterns even though none of the mask fields are perfect.[6.1.2] This method of recovering pattern integrity is especially valuable in situations where quick fabrication turnaround is the primary concern or the means for mask inspection and repair are not available. Fig. 14 illustrates the technique. The top portion of the diagram shows the wafer resist pattern resulting from exposure with one of three reticle fields which contains random defects, the bottom half shows an equivalent pattern resulting from the exposure with all three reticle fields.

A reticle for an Ultratech stepper was generated using MEBES, and contained electrical defect test structures to verify that the voting scheme can eliminate random mask defects in large-area high-density patterning applications. As shown in fig. 15, in one of the three fields used for these experiments, deliberate defects of various sizes were introduced into half of the patterns by inserting or removing rectangles in the original graphics file; this field will be referred to as field C. Both opaque and clear defects were introduced with dimensions ranging from 2μm x 1μm up to 20μm x 20μm. The test structure consisted of interdigitated combs with interleaved serpentines to detect shorts and opens, and had a 2μm nominal line/space width and an area of 2.4mm^2. Half of this field had perfect patterns to serve as control. The other two fields, termed fields A and B, had only perfect patterns. A fourth field provided appropriate alignment targets to facilitate the experiment.

Three sets of experiments were performed as described in the following. Except for the exposure method, all wafers underwent an identical process sequence. Three wafers were used in each group:

Group 1: Each wafer was exposed with a single exposure of field C. This was the control experiment to determine the baseline yield under normal process conditions.

Group 2: Each wafer was exposed with three exposures of field C. This was done to determine the effect of the triplicated exposure scheme on the yield in areas free of defects.

Group 3: Each wafer was exposed in a full voting scheme using all three fields. This was done to determine the impact of vote-taking exposure on the yield in areas affected by defects, through yield statistics for the defect-bearing structures.

Two special processing techniques were employed in these defect studies to minimize the process complexity and hence the number of variables involved, with the aim of obtaining the best estimate of lithographic defect density. One is the use of latent images in resist for alignment rather than fabricating targets into the substrate. The wafers were first exposed with the field containing target features which produced latent target image that could be detected by the automatic alignment mechanism of the Ultratech stepper. The wafers could then be fed back directly into the stepper to be exposed with the pattern fields and, as a result, only one develop step was needed. The other technique consisted of the use of alkaline developer as an agent for consecutive developing and etching as described above. Due to the continuity between developing and etching it can be claimed that the resulting defects are solely lithographic defects.

To determine the yield, 176 and 352 points per wafer were measured for each defect structure and each non-defect structure, respectively. Because no dependence of the results on the defect size was apparent, in agreement with the observation noted in the previous section, all data points associated with the same defect polarity were combined in the calculation.

Examination of the "combs" column of table 2 shows that zero yield was obtained in defect structures in both groups 1 and 2 which used only field C, as expected. The third experiment, however, showed almost perfectly recovered yields in the same structures, indicating nearly complete elimination of the mask-induced defects by the vote-taking exposure scheme. In addition, the yield in the non-defect structures in group 2 was found to be not less than that obtained in group 1, suggesting that the increase in process complexity due to the triplicated exposure has not created any adverse effects. The yield of the serpentines shows 100% yield for both groups one and two. This implies that opens are not a significant contributor to yield reduction at this linewidth. The group three serpentine yield is only 93%. This "low" yield can be explained by the fact that the linewidth for groups one and two was approximately 1.42μm, while the linewidth for group three was approximately 1.1μm. The influence of linewidth on line-opens at the latter linewidth is clearly important. The combs show almost identical yield for both defect-free and defect-bearing structures, indicating full recovery of structures with introduced defects. The serpentines show similar results with almost complete recovery of the defect-bearing structures to the defect-free level.

Conclusions

A novel etching technique for monitoring of photolithographic defects in VLSI processing has been proposed. It has the ability to separate the normally coupled contributions of lithography and etching, to defect density calculations. This is accomplished by accelerating the etch rate of the metal patterns underlying the resist, in the developer solution, by an order of magnitude. To achieve this mode of operation, the developer is altered to be electrolytic after the completion of the develop cycle. The availability of end-point detection in this wet-etching technique is significant. An application of this technique to verifying the benefits of voting lithography is reported.

6.1.6 References

[6.1.1] McCarthy, A. M., Lukaszek, W., Meindl, J. D., "A New Electrical Test Structure Ion Implant Monitor", *Proc. of the IEEE VLSI Workshop on test structures*, Long Beach, California, February 1986.

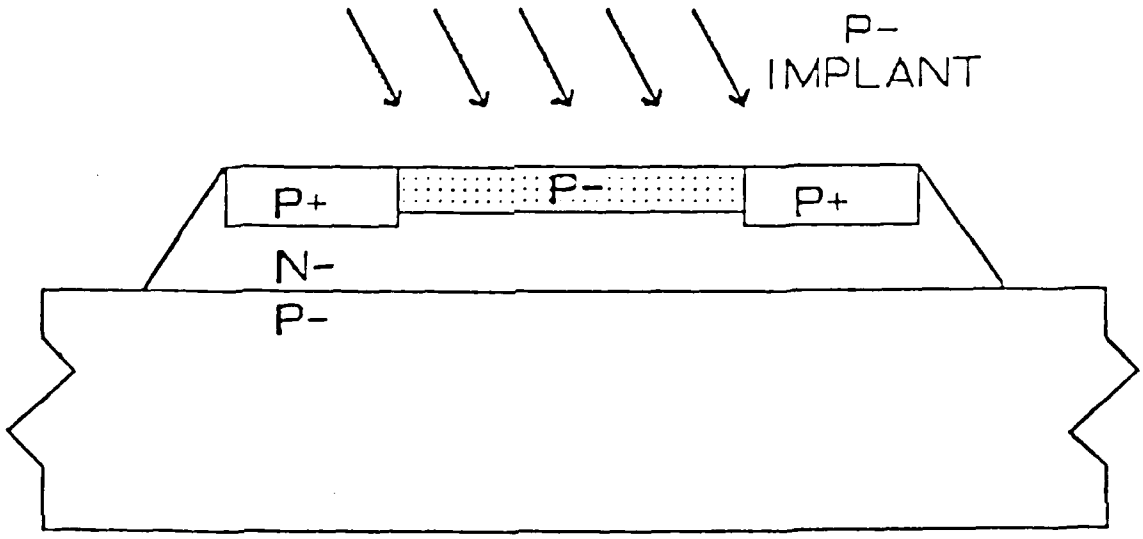
[6.1.2] Fu, C.-C., Dameron, D. H., "Improvement of Mask Limited Yield with a vote-taking Lithographic scheme", *Electron Device Letters*, Volume EDL-5, Number 10, October 1984, pp. 398-400.

[6.1.3] Fu, C.-C., Dameron, D. H., McCarthy, A. M., "Elimination of mask-induced defects with vote-taking lithography", *Optical Lithography V*, *SPIE Journal*, Vol. 633, p. 270.

[6.1.4] McCarthy, A. M., Lukaszek, W., Meindl, J. D., "A novel technique for the consecutive pattern developing and etching of thin aluminum films", To be published. This paper is appended.

[6.1.5] Turner, N. L., Current, M., Smith, T. C., Crane, D., "Effects of planar channeling using modern ion-implantation equipment", *Solid State Technology*, February 1985, pp. 163-172.

FIGURE 1 Plan and cross-section views of ion implant monitor structure designed for dosimetry measurements.



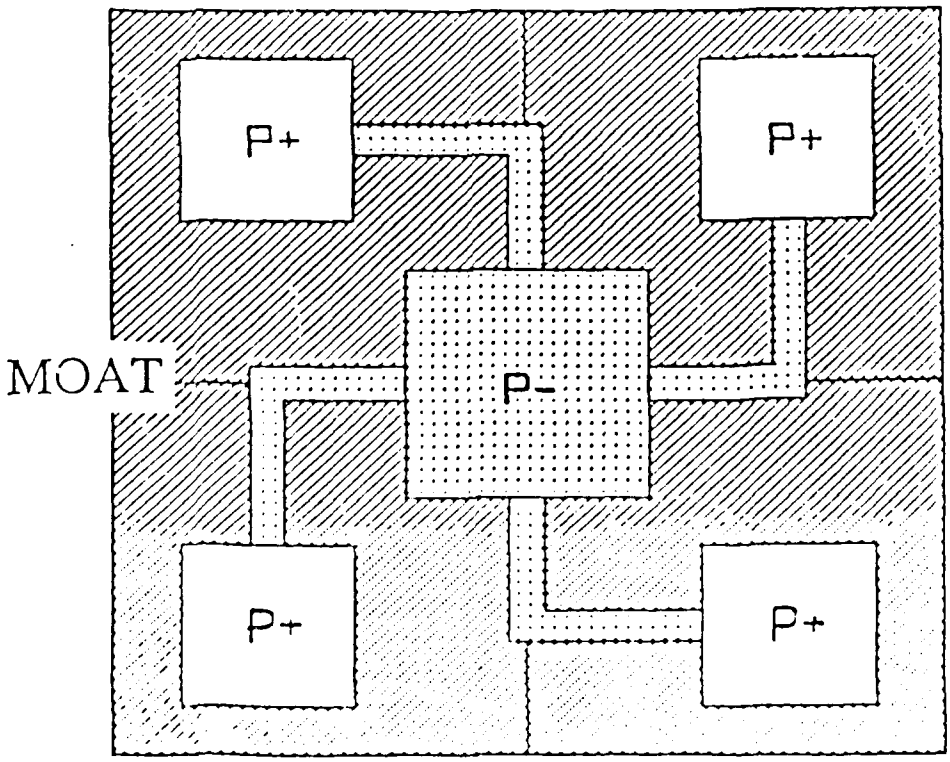


FIGURE 2 Plan and cross-section views of ion implant monitor structure designed for shadowing measurements.

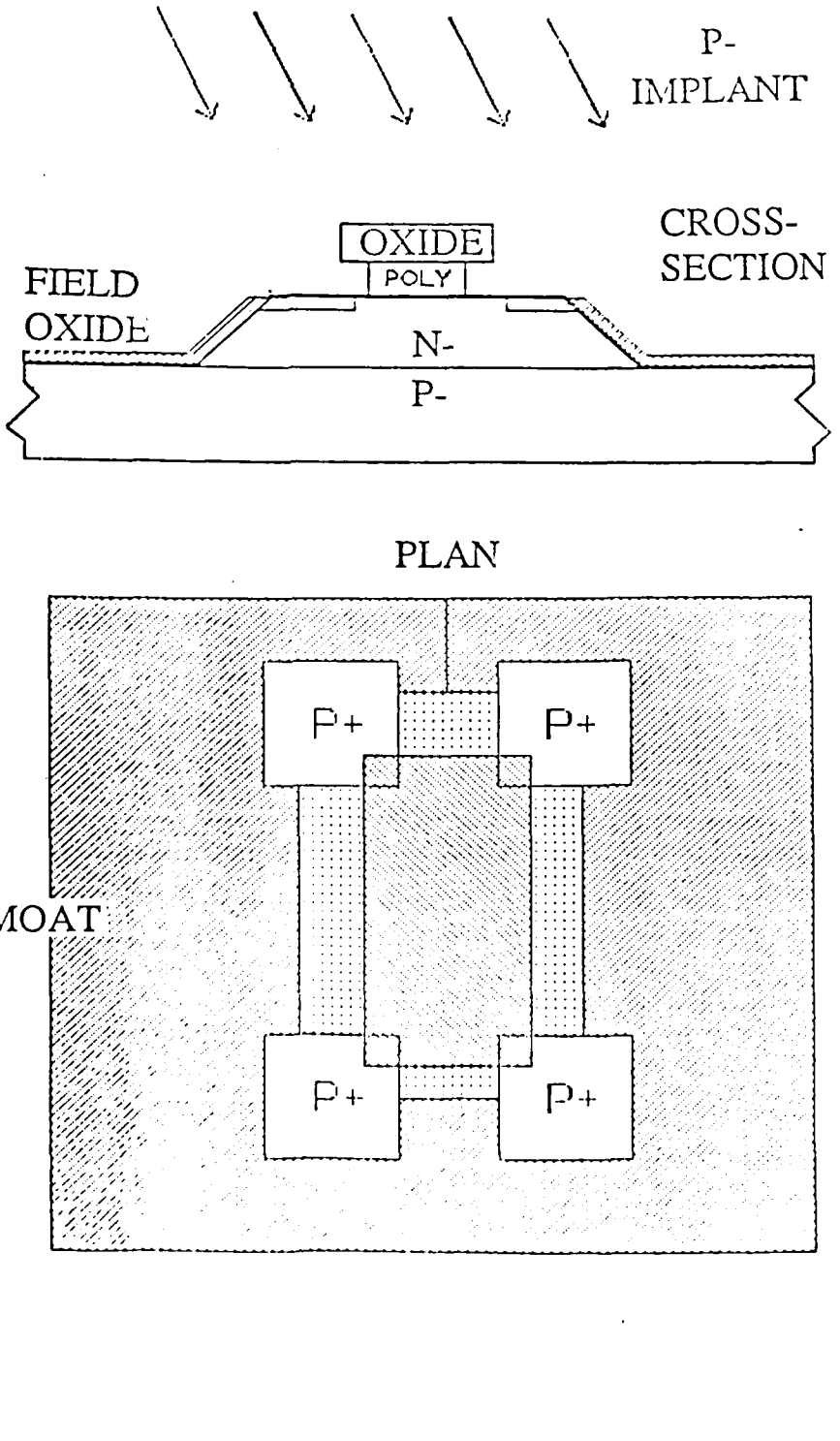
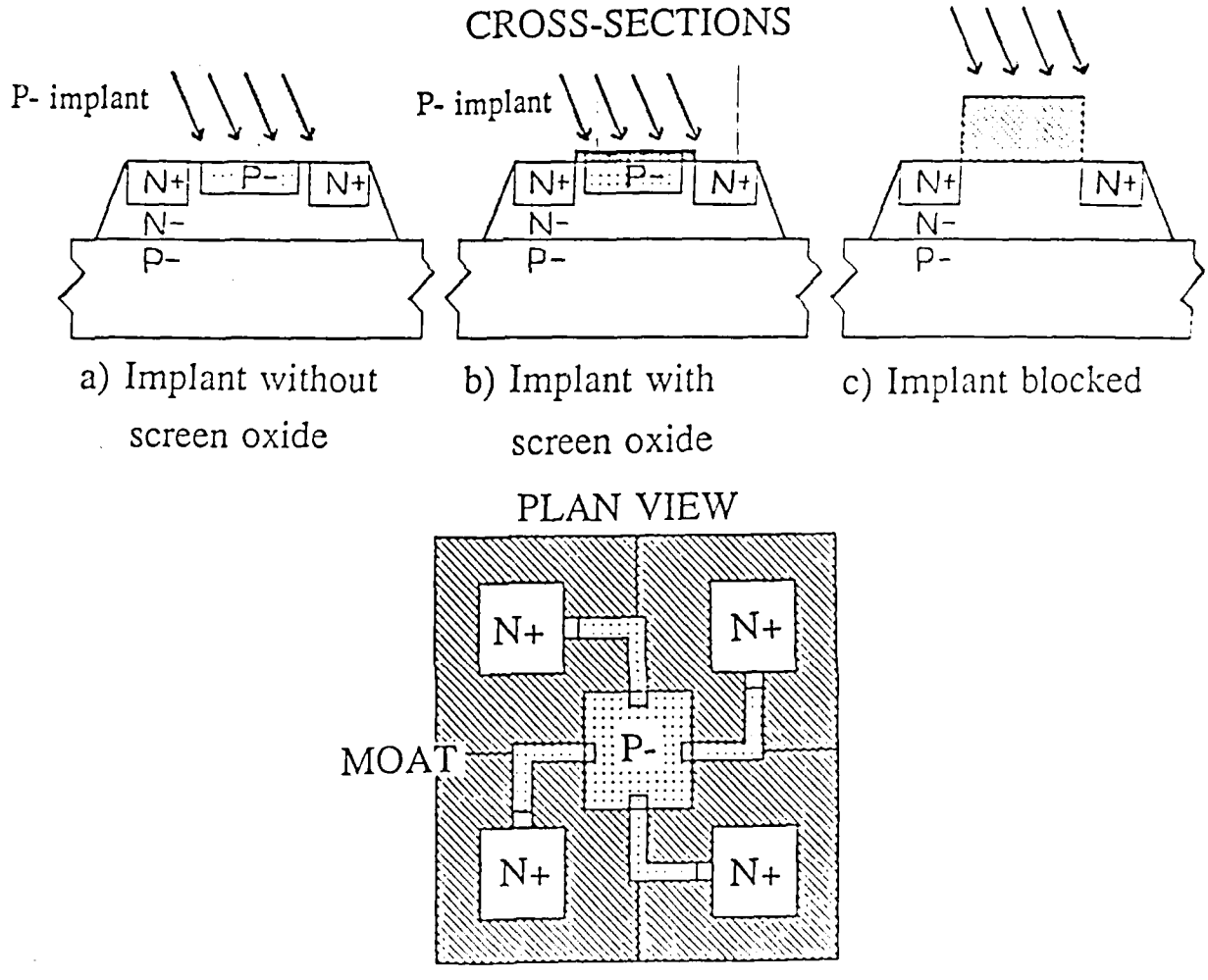


FIGURE 3 Plan and cross-section views of ion implant monitor structure designed for channeling measurements.



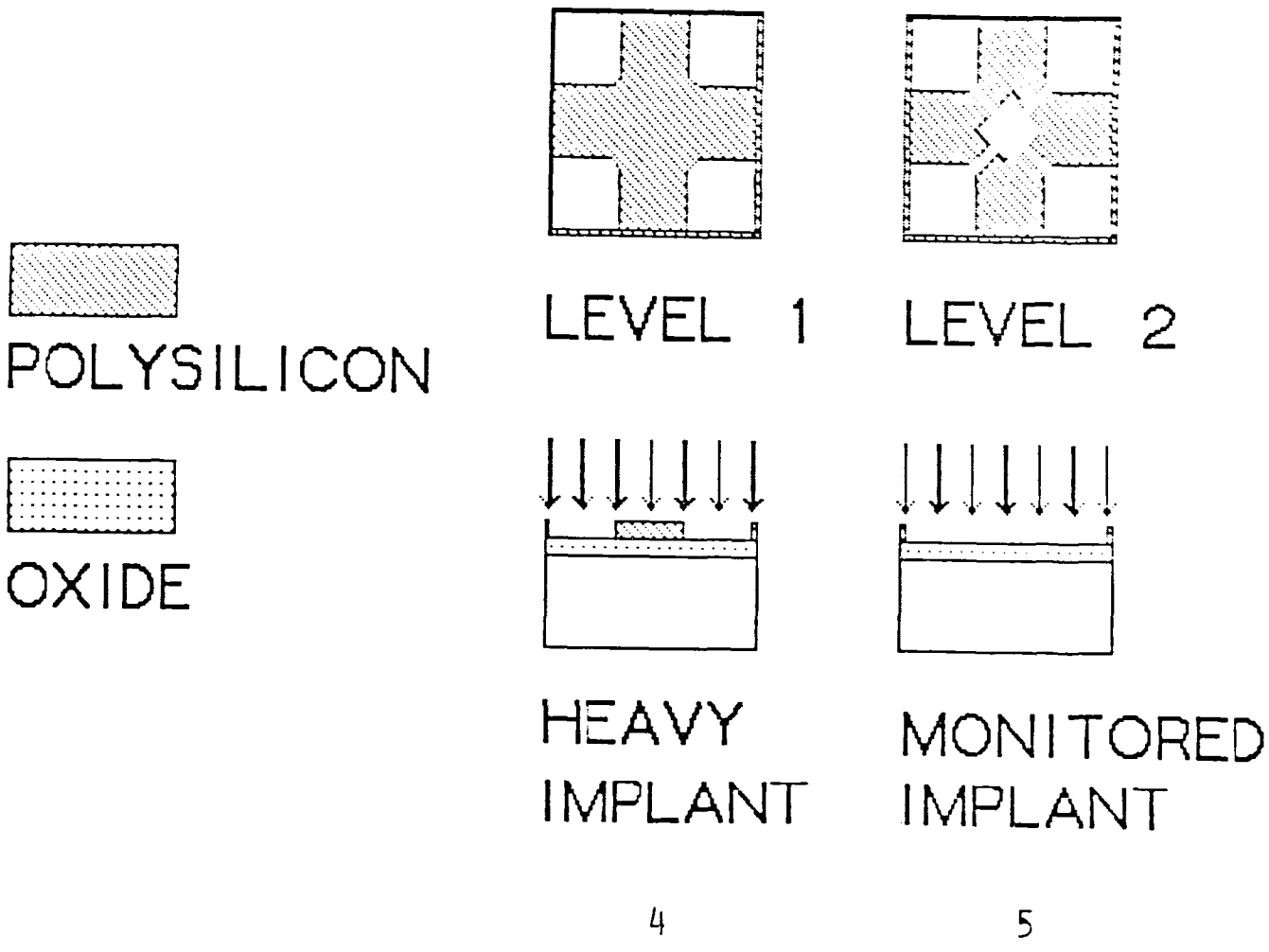


FIGURE 4,5 Plan view of masks,
(a) level 1, (b) level 2.
Cross sections of device structure,
(c) after level 1 processing,
(d) after level 2 processing.

FIGURE 6 Sheet resistance wafer map, boron dose 1e13.

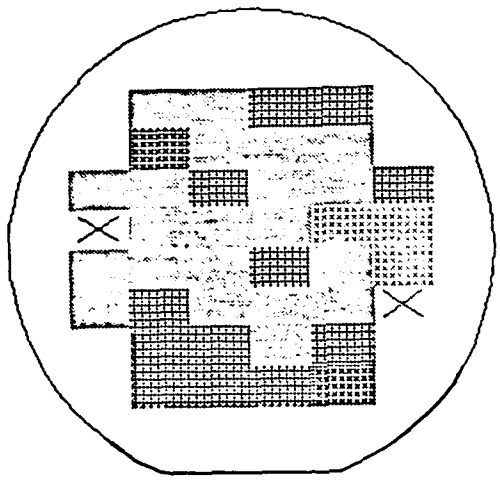


FIGURE 7 Sheet resistance wafer map, boron dose 1e12.

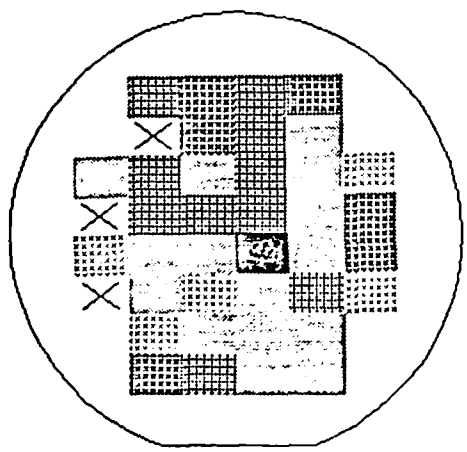


FIGURE 8 Sheet resistance wafer map, boron dose 7e11.

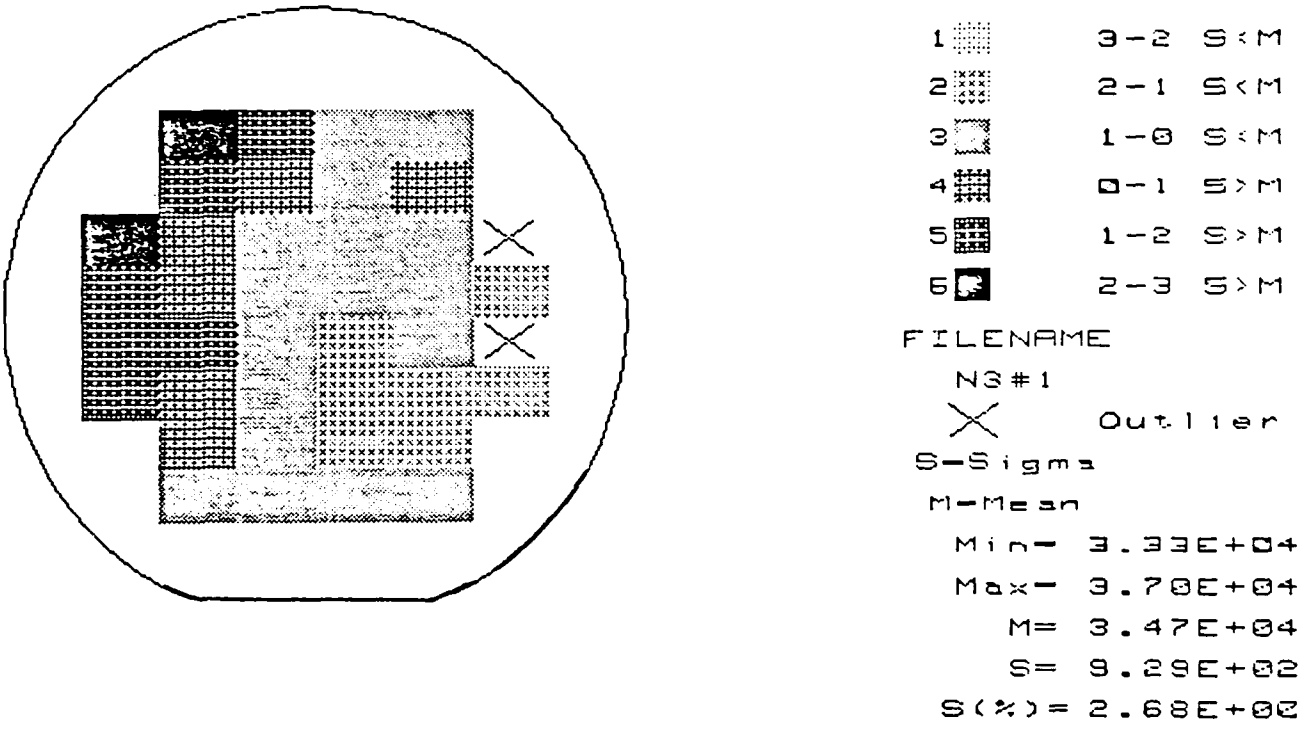


TABLE 1

Within chip measurements						
Dose	Wafer Upper Left	St. Dev.	Wafer Center	St. Dev.	Wafer Lower Right	St. Dev.
ions/cm^2	Ω□	%	Ω□	%	Ω□	%
1e15	126.0	10.0	115.0	5.0	114.0	3.0
	kΩ□		kΩ□		kΩ□	
1e13*	3.1	1.5	3.1	0.8	3.1	0.7
1e12	28.9	1.7	27.2	1.1	27.5	0.8
7e11	42.1	1.1	40.0	1.2	40.5	1.8

* 40mm diagonal on a three inch wafer

FIGURE 9 Dose versus Sheet Resistance

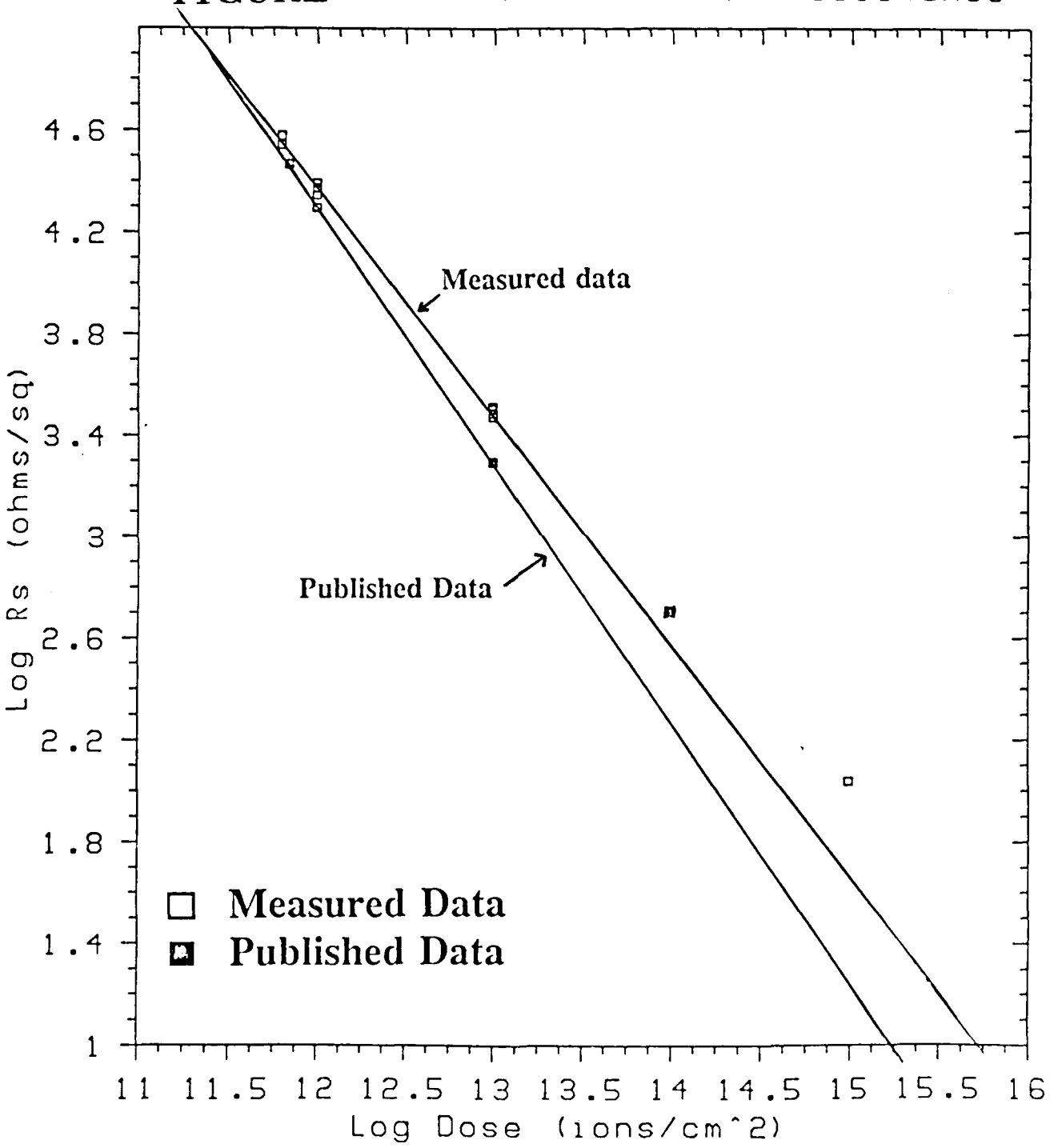


FIGURE 10 IN-PROCESS DEFECT MONITOR STRUCTURE CROSS-SECTION

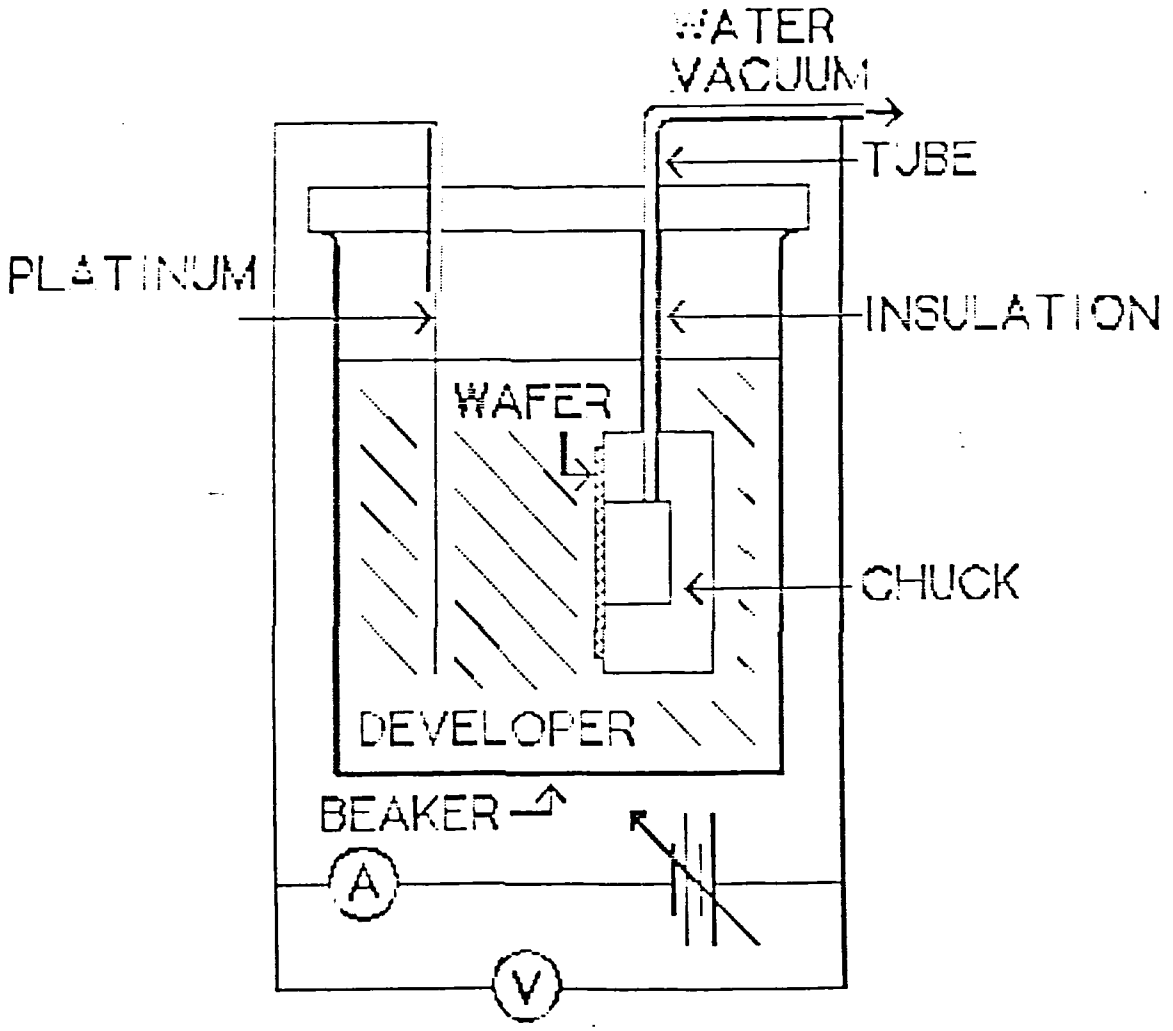


FIGURE 11 IN-PROCESS DEFECT MONITOR EXPERIMENTAL SETUP

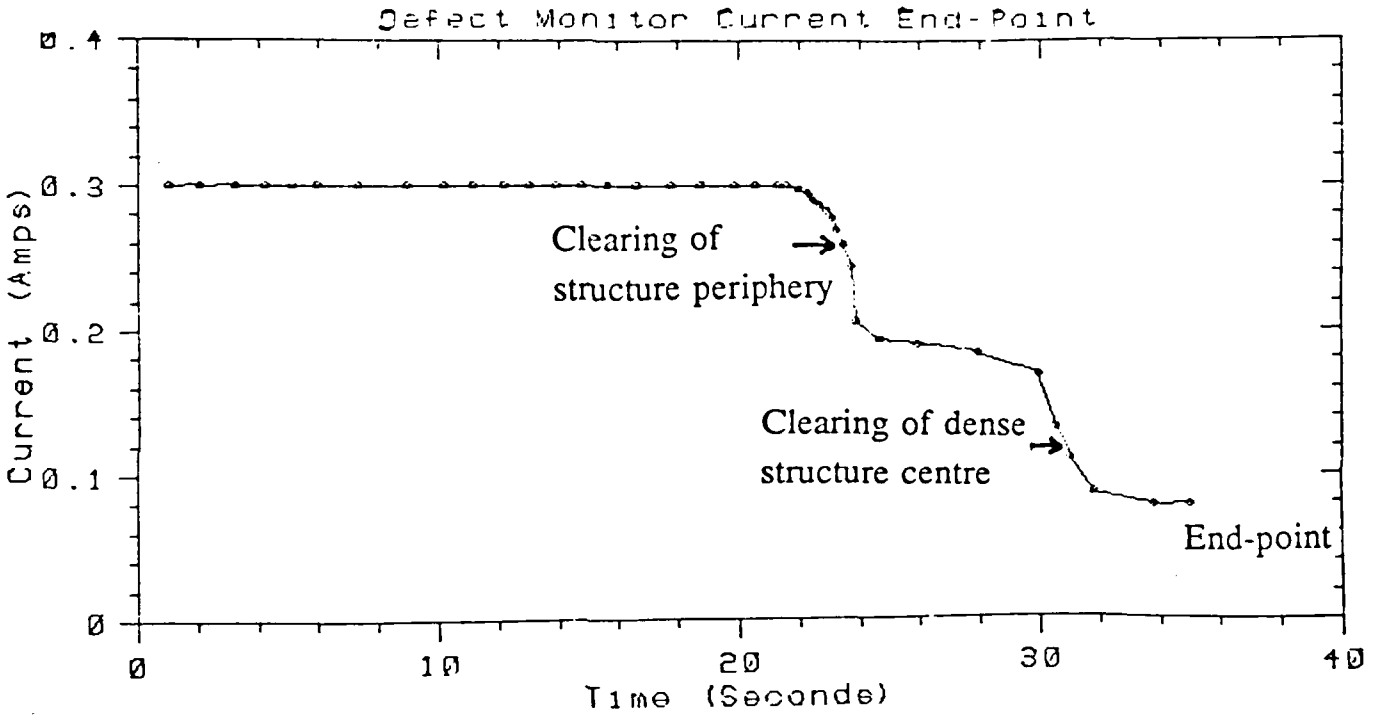


FIGURE 12 Defect Monitor I-V
Electrical End-Point detection

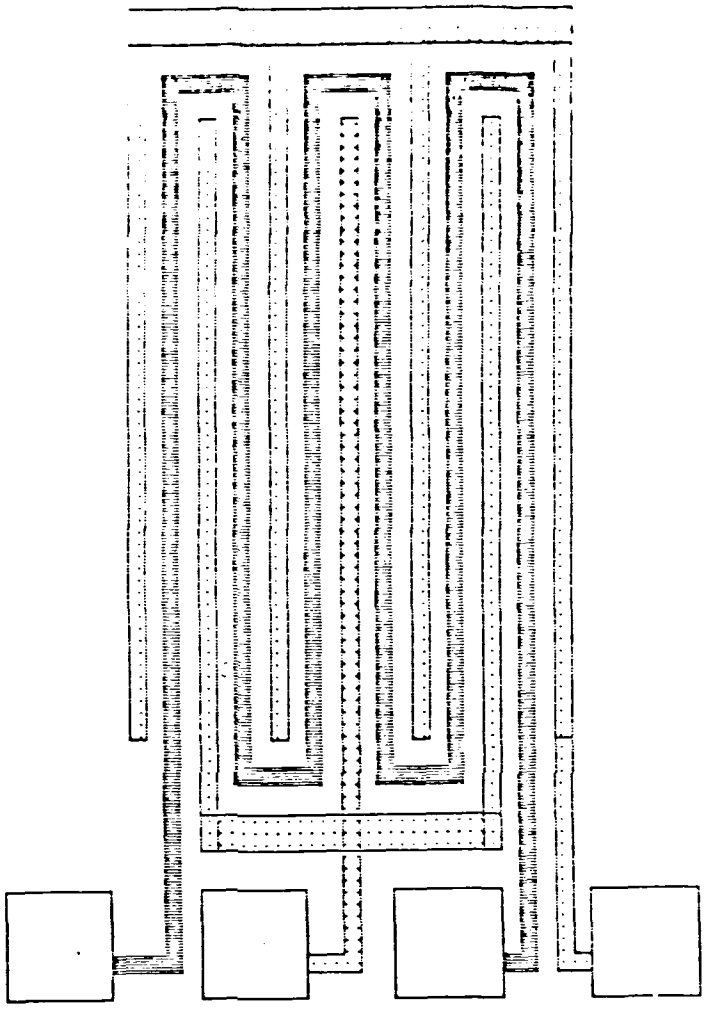


FIGURE 13 Comb/Serpentine

Electrical Test Structure

FIGURE 14 Vote-taking Lithography Scheme

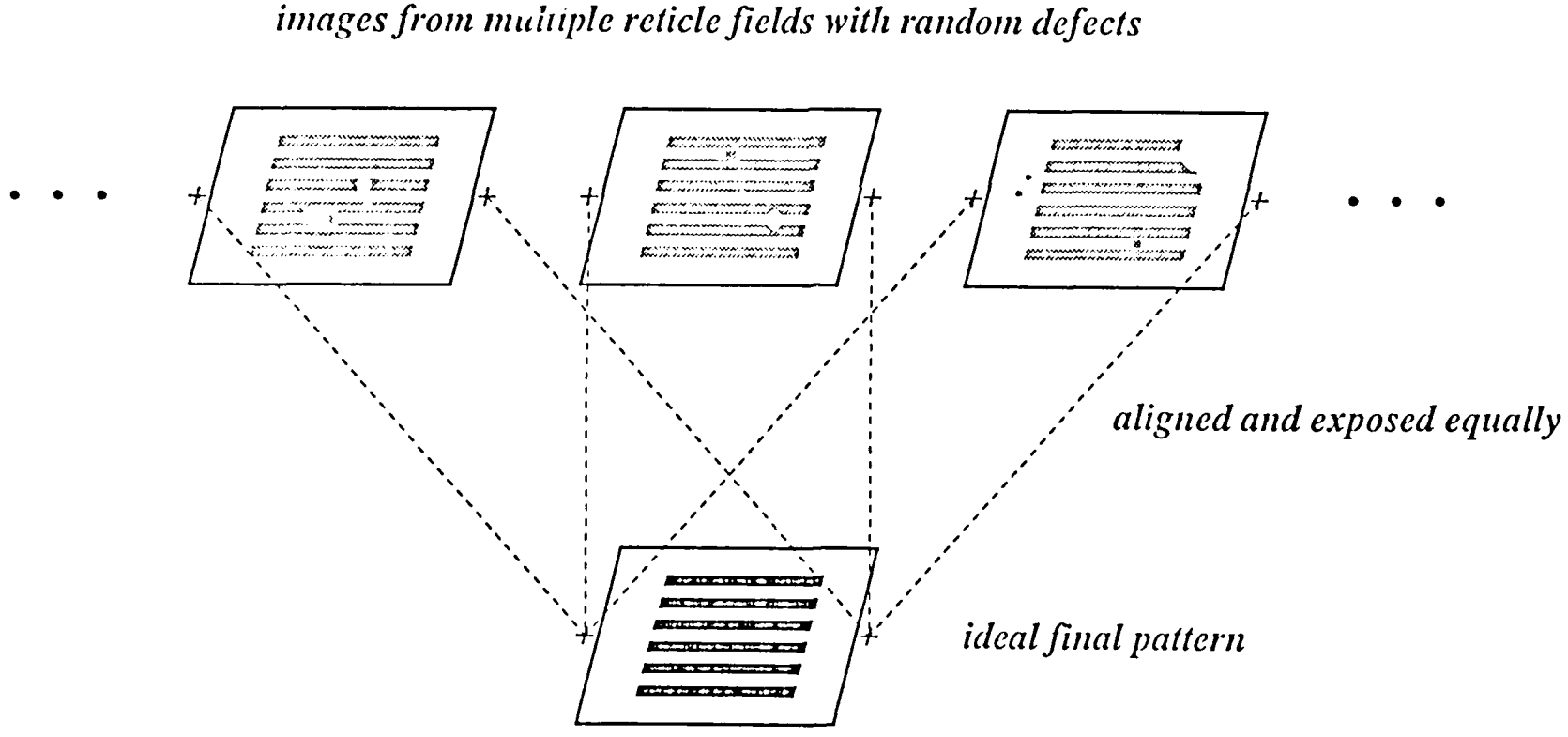


FIGURE 15 Exposure Schemes

Experiment		Measured yield (%)	
		combs	serpentines
Group 1 one field single exposure	defect-free structures	75	100
	defect-bearing structures	0	0
Group 2 one field triple exposure	defect-free structures	83	100
	defect-bearing structures	0	0
Group 3 three fields voting exposure	defect-free structures	90	93
	defect-bearing structures	89	90

TABLE 2 Summary of yield data for combs and serpentines for both defect-bearing and defect-free structures.

6.2 End-of-Process Monitors

W. Yarbrough and W. Lukaszek

The end-of-process monitors consist of thirty two 2x2mm modules arranged in an 8x16mm reticle field [8.2.1] [8.2.2]. Each 2x2mm module consists of one or more unique test structures which in the case of defect density monitors are further divided into geometrically ratioed sections. The parametric module contains several tens of small area structures, including individual transistors, line width monitors, etc. The choice of the defect density monitors was guided by the desire to electrically determine all bulk, interface, and topography related defect densities of our two layer metal, n-well CMOS process. Two general sets of structures are employed for this purpose. One set of structures divides the composite process structures (e. g. transistors) into interrelated sets of simplest possible sub-structures to obtain "elemental" defect densities associated with these sub-structures. Another set of structures carefully combines elements of the simplest structures into more complex structures to examine the additive relationships between different elemental defect densities as a prelude to yield prediction from elemental defect densities. This approach is culminated in modules consisting of 160x160 micron ring oscillator structures which will be used as a test vehicle for studying defect clustering and as a final check on the validity of the yield prediction relationships.

The salient features of the 32 end-of-process test structure modules are described below.

1. GATE OXIDE MODULE 1

This module contains four test structures for evaluating defects in gate oxides. There are two structures, identical in design but different in size, which evaluate the area gate oxides in NMOS and PMOS devices. Two other structures, one for NMOS and one for PMOS, evaluate the gate oxide integrity along field oxide edge. These structures will also be used to evaluate inversion layer-inversion layer isolation. The sizes of these structures represent approximately 85 percent of the area and edge length that occur in the MIPS processor chip, our most ambitious IC.

2. GATE OXIDE MODULE 2

This module contains two test structures, one for PMOS and the other for NMOS, to evaluate the gate oxide integrity along poly edge. These structures will also be used to evaluate source-drain leakage along the channel, and the junction leakage component of source/drain junctions along poly edge. These structures represent approximately 85 percent of the physical MOSFET edge length in the MIPS processor chip.

3. DRAIN-TO-SOURCE LEAKAGE ALONG FIELD OXIDE EDGE MODULE

This module contains two structures, one NMOS and one PMOS, to monitor drain to source leakage along the field oxide edge. They are arrays of transistors, of minimum length and width, connected in parallel by common diffusion busses to minimize the influence of contacts on junction leakage. The number of transistors in these arrays represent approximately 25 percent of the total number of transistors in the MIPS processor chip.

4. METAL TO N+ DIFFUSED REGION CONTACT JUNCTION MODULE

This module contains two structures to monitor area and contact induced leakage of diffused N+ junctions. One of the structures has a large number of contacts, while the other structure has a very small number of contacts. The structure with large diffused area and small number of contacts represents approximately 25 percent of the N+ diffused area that occurs in the MIPS processor chip. The structure with a large number of contacts has 4 connected subsections which have 1, 2, 4, and 8 times the minimum design rule spacing between contacts. This structures represents approximately 20 percent of the number of contacts (and associated metal area) in the MIPS processor chip. All contact windows are 2 micron X 2 micron. Successive rows of contacts are offset to facilitate cleavage through contacts for SEM examinations.

5. METAL TO P+ DIFFUSED REGION CONTACT JUNCTION MODULE

This module is identical to the Metal to N+ Diffused Region Contact Junction Module except that the contacts are to P+ diffused regions.

6. METAL TO N+ DIFFUSION CONTACT MODULE

This module contains two identical structures for evaluating yield of contacts between first level metal and N+ diffused regions. Both structures consists of three separate contact chains which are tapped at several places. These two structures contain approximately 25 percent of the number of metal to N+ diffused region contacts in the MIPS processor chip.

7. METAL TO P+ DIFFUSION CONTACT MODULE

This module is identical to the Metal to N+ Diffusion Contact Module except that the contacts are to P+ diffused regions. The structures in the module contain approximately 31 percent of the number of metal to P+ diffusion contacts in the MIPS processor chip.

8. METAL TO POLY CONTACT MODULE

This module contains two identical structures for evaluating the yield of contacts between first level metal and polysilicon. Each structure consists of one continuous contact chain that is tapped at several places. These structures contain approximately 25 percent of the first level metal to poly contacts in the MIPS processor chip.

9. VIA STRINGS AND METAL COMPOSITE SERPENTINE MODULE

This module contains two structures for evaluating the resistance and yield of vias between first and second level metal. The yield monitor structure consists of a tapped via chain. The other structure consists of several individual via chains of identical lengths and varying numbers of vias to determine the influence of vias on chain resistance. This module contains approximately 25 percent of the number of vias in the MIPS processor chip.

The metal composite serpentine evaluates the metallization integrity of the complete metallization system. This includes metal 2, metal 1, metal 2 to metal 1 vias, metal 1 to diffusion contacts and metal 1 to poly contacts. It will be used primarily for yield prediction exercises.

10. PARAMETRIC STRUCTURES MODULES

This group of structures is used to collect parametric data for process monitoring and device and circuit analysis. The modules include structures for measuring sheet resistance and line width of metal 1, metal 2, poly, N+ diffusion, P+ diffusion and N-well diffusion. Also included are structures for measuring metal 1/N+ diffusion, metal 1/P+ diffusion, metal 1/poly and metal 1/metal 2

contact resistance. 17 N-channel and 17 P-channel transistors are included for SPICE and TECAP parameter extraction. Field transistors are included for field threshold voltage data and for device to device isolation integrity analysis. Inverters are included for inverter characteristics extraction. A 127 stage ring oscillator with a Schmitt trigger starter circuit is included for extracting gate delay and power consumption data. 3 N-type and 3 P-type capacitance structures are included for measuring bottom junction, sidewall junction and gate overlap capacitances for SPICE circuit simulation. CMOS latch-up characteristics are evaluated as function of critical spacings between diffused regions of typical CMOS inverters.

11. Field Isolation Module

The field isolation module contains two structures, one NMOS and one PMOS, which evaluate the integrity of the isolation between diffused regions. They can also be used to monitor metal field threshold voltage. These structures represent approximately 25 percent of the diffused region area, at minimum design rule width, of the MIPS processor chip.

12. Metallization Step Coverage and Photolithographic Proximity Effects Structures Module

The step coverage structures are used to evaluate metallization shadowing and step coverage over increasingly difficult topography created by the underlying layers. Each structure isolates a particular topographical situation which may aggravate lithography and/or step coverage.

The photolithographic proximity effects structures are used to detect systematic problems in metal lithography due to reflections which unintentionally expose resist patterns in adjacent valleys. Two types of patterns are used in each of the 3 photolithographic proximity effects monitors. One pattern consists of metal lines running parallel to metal and/or poly lines. The other pattern consists of metal lines running over a metal and/or poly "waffle pattern" which simulates the worst case photolithographic topography that metallization could encounter.

13,14. Composite Metallization Array (Comb) Modules

These modules are used to evaluate metallization intralevel and interlevel shorts on structures emulating the cache memory of the MIPS-X chip. The structures contain equivalent amounts of diffusion, poly, metal 1 and metal 2 areas and edge lengths that occur in the cache memory. The amounts of interlayer overlap and intralayer spacing are also equivalent to that in the cache memory. These structures will also be used for metalization yield prediction from defect densities data obtained on metalization decomposition structures.

15,16. Composite Metallization Array (Serpentine) Modules

These modules are used to evaluate metallization opens and interlevel shorts. The structure has the same unique design as the Composite Metallization Array (Comb) structure, described above. These structures will also be used for metalization yield prediction from defect densities data obtained on metalization decomposition structures.

17,24. Metallization Decomposition Structures Modules

These structures are used to evaluate all aspects of the two level metallization system. They differ from modules 11-14 in that they isolate one aspect of the two level metallization at a time. Consequently, there are 8 structures which evaluate interlayer and intralayer isolation, step coverage, and lithography. All of these structures are used strictly for problem decomposition analysis

and elemental defect densities extraction and do not attempt to emulate any aspects of the MIPS-X cache memory.

25-28. N AND P-CHANNEL TRANSISTOR ARRAYS MODULES

These modules evaluate gate oxide integrity, source/drain junction integrity, drain to source isolation integrity, and device to device isolation in one composite structure. Therefore, they are a recomposition of all the decomposition test structures described above, and can be used to verify that the results of all the decomposition test structures, taken collectively, are accurate. They will be used primarily for yield prediction exercises.

29,32. RING OSCILLATOR ARRAYS MODULES

The yield data obtained from these arrays will be used to verify the feasibility of yield prediction using component defect density data obtained from the previously described process partitioning test structures. Defect clustering information obtained from these structures will also be used to evaluate the assumptions underlying the commonly accepted, and typically inadequate, yield formulae, and to examine the feasibility of manufacture of daringly large ICs.

Supporting Activities

In order to establish a reliable testing facility to support the above activities and to facilitate the transfer of the research findings to industry, the test group has also accomplished the following during this reporting period:

1. COLLABORATIVE TESTING AND EVALUATION

The test group recently established a collaborative effort of testing and evaluation of the previously described in-process and end-of-process monitors with colleagues at Hewlett-Packard Labs. The industrial mentor overseeing this activity is Dr. Dirk Bartelink. The intent of this effort is to jointly develop the appropriate test and data analysis software to assess the performance of the above mentioned monitors, and to accelerate their transfer to industry.

2. INSTALLATION OF PARAMETRIC TEST SYSTEM

The collaborative testing activity with Hewlett-Packard Labs will be greatly facilitated by HP's donation to CIS of a complete HP4062B parametric test system, including a HP9836CT control computer, a HP7946A 55 Mbyte Disc/Tape Drive, a HP2671G graphics printer, and a HP7475A graphics plotter. A donation by Rucker and Kolls of their 1032 auto-stepping prober completes the system, and duplicates the test installation at Hewlett-Packard Labs. This will assure compatibility of jointly developed software and facilitate the reciprocal transfer of software and data between the two installations.

3. DONATION OF ENHANSYS DATA ANALYSIS SOFTWARE

The testing group has received a donation of a statistical analysis software package from ENHANSYS, Inc. The ENHANSYS software, also employed at Hewlett-Packard Labs, will be used for statistical analysis and reduction of the large amounts of data obtainable from our end- of-process monitors.

4. Donation of RS/1 Data Analysis Software

The testing group has received a donation of a statistical analysis software package from Bolt, Beranek and Newman, Inc. (BBN). This software will be used in addition to the ENHANSYS software. Its broad applications base allows it to be very flexible and it will be used for statistical analysis which will supplement the ENHANSYS software capabilities.

Wafer Processing and Testing

1. Wafer Processing

Several CMOS wafer lots containing modules 1-16 have been processed and preliminary evaluation of wafers from each lot indicate that the design and layout of the structures is correct. Additional lots are in progress and are scheduled for completion in October, 1986.

2. Test Software Development

Test software for measurement of all test structures has been written and debugged. Additional refinements of the software is in progress.

3. Test Measurement Results

Preliminary testing has been done on wafers from several CMOS lots. Initial analysis show that most of the test structures are producing expected results. However, some test structures revealed problems in our process and have played an extremely vital role in debugging and correcting these process problems. Additional work is underway to refine data acquisition techniques and to establish appropriate measurement conditions.

4. Test Data Analysis

Two statistical software packages have been installed for use in data analysis of the measurement data. At present, we have begun the learning process for using these software packages and exploiting their capabilities. Additional work is needed to determine the appropriate data presentation methods. This work is in progress.

REFERENCES

[8.2.1] Yarbrough, W., Lukaszek, W., and Meindl, J., "VLSI Process Problem Diagnosis and Yield Prediction: A Comprehensive Test Structure and Test Chip Design Methodology", IEEE VLSI Workshop on Test Structures, Feb. 17-18, 1986, Long Beach, CA, pp. 294-304.

[8.2.2] Lukaszek, W., Yarbrough, W., Walker, T., and Meindl, J., "CMOS Test Chip Design for Process Problem Debugging and Yield Prediction Experiments", Solid State Technology, March 1986.

6.3 Determination of Optimal CMOS Design Rules

Greg Freeman, Tsu-Chang Lee, W. Lukaszek

6.3.1 Introduction

Current industrial practice in determination of layout design rules is often based on judgement, with an inadequate amount of data to back it up. These rules are then fixed until a circuit designer wishes to stretch the limits by incorporating "special case" situations in his design.

Ideally, design rules should be determined by a yield/performance/cost tradeoff in accordance with the goals set forth for the product. Yield can usually be improved by implementing conservative design rules. Performance, on the other hand, is optimized by taking rules to their smallest limits while keeping yield greater than zero. This project is devoted to research to make a system capable of such design rule determination feasible.

6.3.2 System Description

The system consists of three major components, as described below.

1. Test structures to measure alignments, critical dimensions, defect densities, and device properties (latchup, etc.).

2. Data reduction module. Since an enormous amount of data can be collected from test structures, it is best to make it manageable early in the cycle. The output of this module will be descriptions such as means, variances, etc. for each measurement type.

3. Data interpretation module. This module utilizes the reduced data and creates a listing of the design rules. It may contain a system similar to that described in [6.3.1], called STRUDEL.

Because the data reduction and interpretation components of the system depend on which test structures are implemented, the design of a complete set of test structures should be the first step in the development of a system. This document describes the progress made to date in the test structure design effort.

6.3.3 Test Structures

There are two primary goals in designing this set of test structures. One is to create a complete set of structures. In other words, we wish to obtain all the information we require for design rule determination from test structures. The only exceptions will be that information which can be accurately simulated. The other goal is to obtain the most accurate set of structures possible. Accuracy is crucial to the success of the system because design rules usually push the equipment to its limit, and this allows little room for uncertainty in measurement. Fulfillment of both goals is a necessary requirement for a fully analytical method of design rule determination.

In designing these test structures, one notable layout restriction applies. Because the mask exposure system writes in blocks 128 microns high and 1024 microns wide, there exist discontinuities

in the masks at the boundaries of these areas. Since we are aiming for maximum accuracy, these discontinuities must be avoided. Fortunately, the pad dimensions of 160 microns between centers gives four pads the same period as five exposure blocks of 128 microns high. This means that if the pairs of test structures are designed to be four pads high, and if the structures are placed top to bottom, the horizontal exposure boundaries will lie in the same place for all test structures. This allows us to design around the boundaries without regard to structure placement.

Alignment Structures

In order to examine the requirements that the alignment structures must fulfill, the Stanford 2 micron CMOS set of masks has been chosen as the work vehicle. The list of masks, in the order that they are applied, is as follows:

1. N-well definition
2. Active field - defines the nitride areas for forming the thick oxide
3. N-well protect - this, in addition to the active field mask, defines where the field threshold implant goes.
4. N-channel implant - this is essentially another n-well protect
5. Polysilicon
6. N+ source/drain - the edges of this mask are the same as the active field minus the P+ s/d regions and bloated 1.5 microns.
7. P+ source/drain - the edges of this mask are the same as the active field minus the N+ s/d regions and bloated 1.5 microns.
8. Contact
9. Metal1
10. Via
11. Metal2
12. Pads

Note that the N+ s/d and P+ s/d masks simply cover the P+ s/d region and N+ s/d regions, respectively, and that their edges lie over the thick oxide. This implies that their misalignments will not be noticeable unless they are very far out of alignment and cover or expose part of the wrong thin oxide region. Therefore, under good alignment conditions, the N+ s/d and P+ s/d edges on the wafer will both be defined by the thick oxide edges (active field mask).

In determining alignment, two possible methods can be used: electrical or visual. Only electrical alignment structures will be considered for use because of the automated nature of the design rule system and the superior accuracy of electrical alignment structures over visual alignment structures. The electrical alignment structures considered here can be classified into four general types. We call these the split resistor, tapped resistor, transistor width and digital vernier structures. Often,

the alignment is best measured by only one type of structure. If the same purpose can be served by more than one structure, the most accurate structure must be determined. In some cases, the most accurate structure will be obvious from theoretical analysis. In others, more than one structure must be designed and fabricated, and the most accurate determined empirically.

For each type of structure, a single misalignment vector will be obtained from two nearly identical test structures: one structure for the x component of the vector and the other structure rotated ninety degrees for the y component. Fig. 1 shows a pair of split resistor structures and a pair of tapped resistor structures. These four structures will give two misalignment vectors. Note that each pair of structures are located as close as possible to each other since misalignments may vary across the chip.

Split Resistor Structure Split resistor alignment structures can be designed if the two masks whose relative alignment is being measured define the opposite edges of a conducting layer. Figs. 2 through 5 are examples of this type of structure. Each of these structures gives the horizontal component of misalignment. Not all of the pads are shown. One half of each structure gives the misalignment measurement, and the other half gives the sheet resistivity which is used in the misalignment calculation. Both halves also give other useful information such as critical dimensions which is not needed for the misalignment measurement.

Operation and Measurements Consider the poly to active area alignment structure of Fig. 2 as an illustrative example. The structure works as follows. A constant current, I, is applied between the top two diffusion contacts and four voltages are measured, two in each of the upper and lower halves of the structure. If the alignment is perfect, then the bottom two voltages should be the same, since the poly will divide the two diffusion strips equally. If not, then the poly will divide the diffusion strips unequally and the voltages will give the misalignment through the formula:

$$\Delta W = \frac{I\rho_s L}{2}\left(\frac{1}{V_1} - \frac{1}{V_2}\right) \tag{6.3.1}$$

where V_1 is the voltage on the side of the structure towards negative poly to active area misalignment (the left, if +x goes to the right), L is the distance between taps, and ρ_s is the sheet resistivity determined from the top half of the structure.

Sheet resistivity can be found from the top of the structure because we have two measurements and two unknowns. The two unknowns are the sheet resistivity, ρ_s, and the difference between the designed and actual widths of each of the lines, E. For example, if the lines are designed to be 3 and 4 microns and are 3.2 and 4.2 microns when fabricated, then E is .2 microns. The formulas giving these parameters are the following:

$$\rho_s = \frac{W_3 - W_4}{IL}\left(\frac{1}{V_3} - \frac{1}{V_4}\right) \tag{6.3.2}$$

$$E = \frac{W_3 - W_4}{V_3}\left(\frac{1}{V_3} - \frac{1}{V_4}\right) - W_{D3} \tag{6.3.3}$$

where W_3 and W_4 are the actual widths of the top two segments and $W_3 - W_4$ is known because it is fixed by the design, regardless of the actual values of W_3 and W_4. W_{D3} represents the designed width corresponding to W_3.

Design The structure must be designed to minimize measurement errors. To do this, three sources of error were identified that may be affected through the design. Consider only the bottom half of the structure, and define L as the length between taps, W as the designed distance between the poly edge and thick oxide, or the designed width of each diffusion strip, and D as the tap width. The design will then minimize the measurement error for each source of error listed below as follows:

1. Finite tap width - the finite width of the tap, D, perturbs the current flow through the long diffusion strip, and decreases the measured resistance: maximize W, L, and minimize D

2. Surface and side non-uniformities - layer thickness and width variations introduce local disturbances: maximize L

3. Machine measurement error - the voltmeter itself has a limit to its accuracy, which suggests we should maximize the fractional change in the measurement: minimize W

Of these, we were able to obtain analytical expressions for (1) and (3).

The effect of finite tap width was analyzed in [6.3.2] through the method of conformal transformations. When the formula (48) from [6.3.2] is applied to determine the error in measuring the misalignment, it was found that the errors from each side of the structure will cancel at zero misalignment, and that they increase with increasing misalignment. However, unless the misalignment is very large, the measurement error remains very small. For a misalignment of .1 microns, $L = 60$ microns, and $D = 3$ microns, the error is less than .01 percent of ΔW as long as W is greater than 3 microns. In addition, all the dimensions will scale.

The other primary source of error, that due to the voltmeter limitations, can be considerably greater than the tap width error. The percent error can be easily derived as being $We/\Delta W$, where e is the percent accuracy of the voltmeter (A typical best accuracy for a digital multimeter is approximately .001%). These results suggest that the design should incorporate the smallest W possible.

In consideration of the three sources of error described above, the following layout guidelines were established:

1. L: maximum possible - for the horizontal alignment structure, L is limited by the 128 micron exposure window size. 110 microns was chosen for the horizontal alignment structure, and 220 microns was chosen for the vertical alignment structure.

2. W: minimum design rule

3. D: minimum design rule

The structures which follow these guidelines should give an error of less than .3 percent error for $\Delta W = .01$ due to the three sources of error listed above.

Tapped Resistor Structure Tapped resistor structures can be designed to measure the alignment between two masks if those masks define two conducting layers: one which is a contact to the other. Figs. 6 through 12 are examples of this type of structure. The structures shown give the vertical component of misalignment, and not all the pads are shown.

Operation and Measurements Their operation is described in [6.3.3]. A current is forced through the length of the structure. Three voltages are then measured, one between the two topmost taps, V_1, one between the second tap and the contact, V_2, and one between the contact and the bottom tap, V_3. The misalignment is then

$$\Delta S = \frac{L}{2}\left(\frac{V_3 - V_2}{V_1}\right) \tag{6.3.4}$$

where ΔS is the misalignment in the vertical direction, L is the distance between the two topmost taps, and positive misalignment of the contact mask to the contacted mask is up.

Design Several issues also must to be confronted in this design in order to minimize measurement error. They are listed below, where a is the width of the contact divided by the width of the conducting strip, and S is the distance from the edge of the contact to the center of the tap. The design will minimize the resistance measurement error for each source of error listed below as follows:

1. Finite tap width: maximize S, W, minimize D
2. Current patterns due to crowding through the contact (for contacts of high conductivity relative to the conductance of the contacted layer): maximize a, S
3. Added resistance due to current crowding: maximize a
4. Machine measurement error: minimize S
5. Surface and side nonuniformities: maximize S

Eqn. (48) from [6.3.2] can be used for analyzing the error due to the finite tap width. The result is that such errors cancel for all values of D, S, W, and misalignment. Therefore, error due to (1) above is zero.

Similarly, the added resistance due to current crowding cancels when the misalignment is calculated. Therefore, the error due to (3) is also zero.

The current pattern in this structure, assuming (worst case) the contact is of much higher conductivity than the contacted layer, can be calculated from eqns. (37) and (38) of [6.3.2]. The results indicate that for nearly all values of a, the current patterns settle before $S = 1.5W$. Contact width, a, has little effect on the current pattern once this point has been passed.

Machine measurement error is affected similarly to the split resistor structure. The percent error is approximately $Se/\Delta S$, where e is again the percent error in the multimeter measurement.

The results indicate the following guidelines for design of the tapped resistor structures:

1. a: minimum design rule
2. W: a + minimum design rule contact surround
3. S: $2W$
4. D: minimum design rule

These guidelines will produce an error due to the machine measurement error of approximately 1 percent in ΔS, assuming a W of 5 microns and a ΔS of .01 microns.

Transistor Width and Digital Vernier Structures Both the transistor width and digital vernier structures are good alternatives in some cases for the split resistor or tapped resistor structures already designed. Both types of structure will be explored in the coming year.

The transistor width structure is a construction of a transistor whose width varies with poly to active field misalignment. The digital vernier structure [6.3.4] is an alternative to measuring the misalignment to metal layers through resistance type measurements, which may not be adequate due to metal's low sheet resistivity.

Complete Set of Alignment Structures The structures shown in Figs. 2 through 12 constitute a complete set of alignment structures for the Stanford 2 micron CMOS process. By a direct measurement, or through several measurements and some vector addition, the relative alignment of one mask to any other can be determined.

With the addition of the transistor width structures and the digital vernier structures, the set of structures should also be the most accurate obtainable.

Critical Dimension Structures

Another set of measurements needed in determining design rules is critical dimensions. Critical dimensions include linewidths, contact sizes, etc. Linewidths for some layers can be determined from the split resistor structures, and the remaining linewidths can be determined from the tapped resistor structures with resistivities from the van der Pauw structures shown in Fig. 13. The remaining critical dimension structures will be designed in the coming year.

Defect Density Structures

Defect densities such as spot density and size are needed to determine optimum design rules. While work on structures with the goal of design rule determination has not been attempted, similar work has been completed on a large set of structures designed for process debugging, which include structures for defect density determination [6.3.5]. It is anticipated that a subset of these structures will be modified to supply the defect density data needed for determination of optimum design rules.

Device Property Structures

Some important device properties will need to be known in determining design rules. For example, breakdown voltage plays an important role in determining the minimum length of a transistor. Latchup properties as a function of source/drain placements will be considered when making the rules for minimum distance from N-well edge. Test structures to give these and other properties will be designed in the coming year.

6.3.4 Future Work

As noted in the above sections, the design of test structures is not complete. We intend to consider each undesigned structure and optimize it for maximum accuracy. What constitutes each design rule will be researched in order to ensure that we eventually have a complete set of test structures.

When test structure design is complete, the two remaining parts of the system will be researched and previous work either improved or added to.

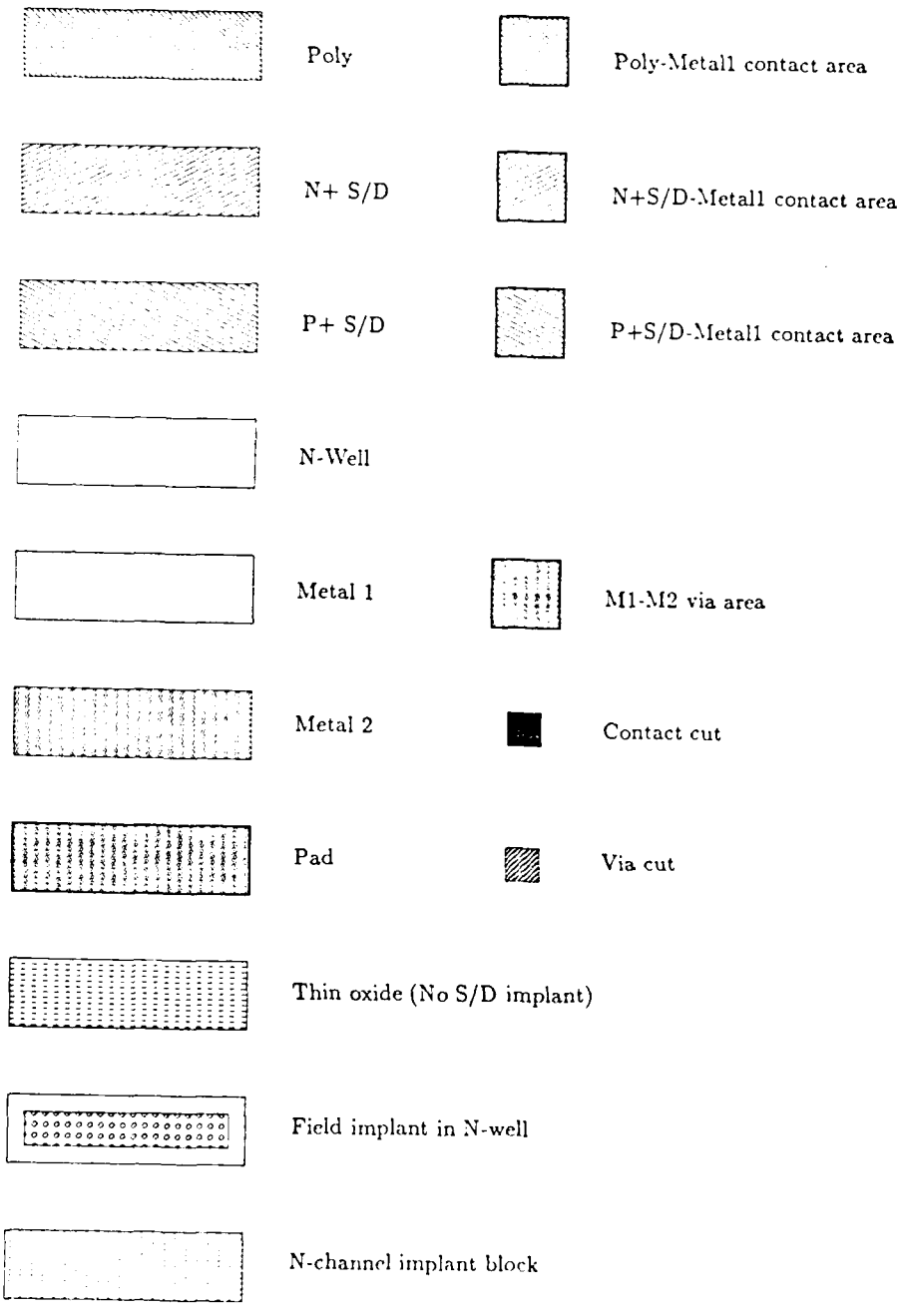
[6.3.1] Rahul Razdan and Strojwas, A., "Statistical Design Rule Developer", *Proc., 1985 IEEE International Conference on CAD* pp. 315-317.

[6.3.2] P. M. Hall, "Resistance Calculations for Thin Film Patterns", *Thin Solid Films*, 1, 1967, pp. 277-295.

[6.3.3] Martin G. Buehler, "The Use of Electrical Test Structure Arrays for Integrated Circuit Process Evaluation", *J. Electrochemical Society*, vol. 127, pp. 2284-2290, Oct. 1980.

[6.3.4] R. Yamaguchi, Komatsu, K., Moriya, S., Harada, K., "Integrated Electrical Vernier to Measure Registration Accuracy", *IEEE Electron Device Letters*, Vol. EDL-7, No. 8, August 1986, pp. 463, 464.

[6.3.5] W. Lukaszek, et. al., "CMOS Test Chip Design for Process Problem Debugging and Yield Prediction Experiments", *Solid State Technology*, March 1986, pp. 87-93.



6.3.1: Layer designations for test structures

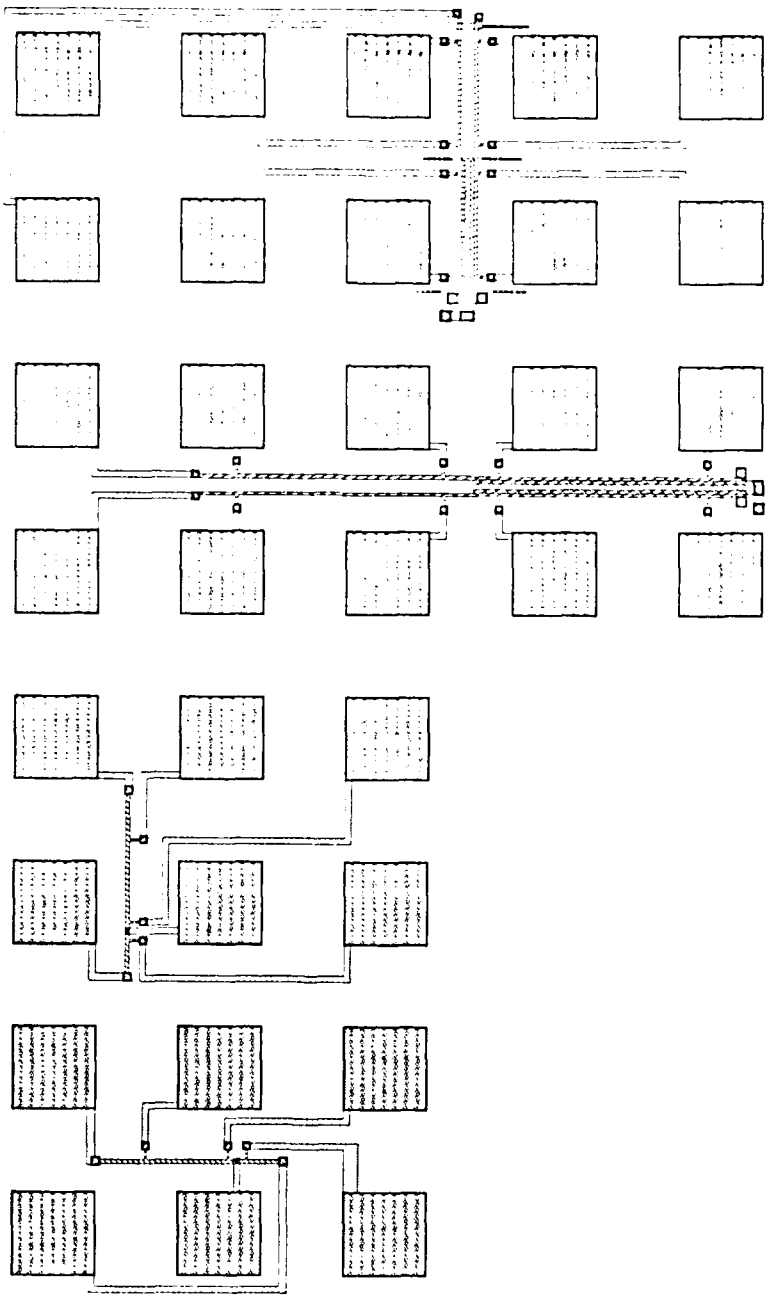


Fig. 1: Four electrical alignment test structures

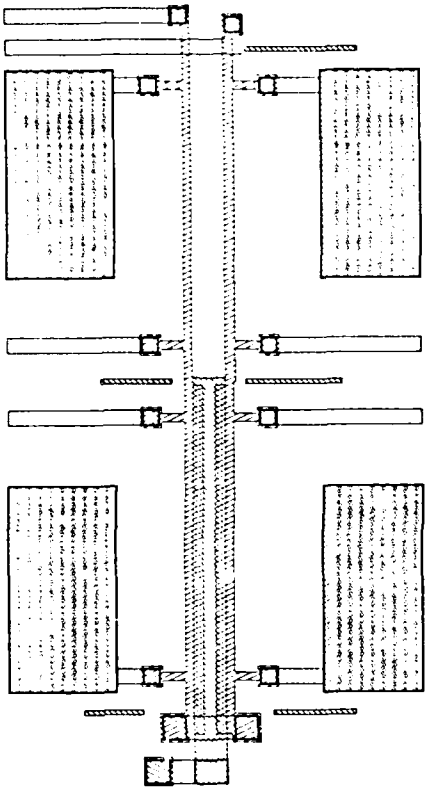


Fig. 2: Poly to Active area

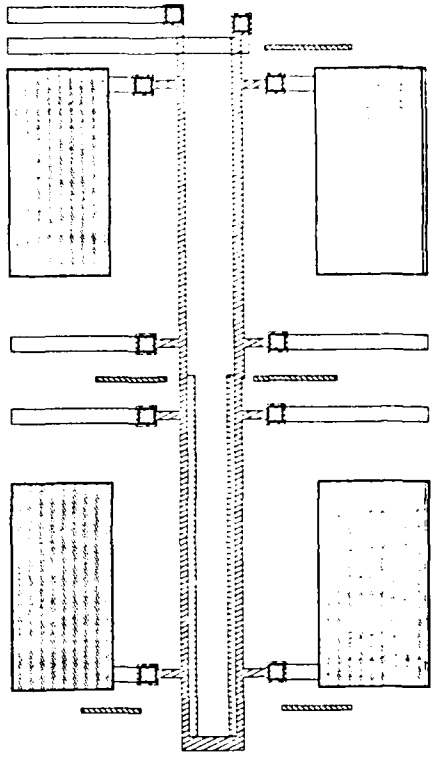


Fig. 3: N+S/D to Active area

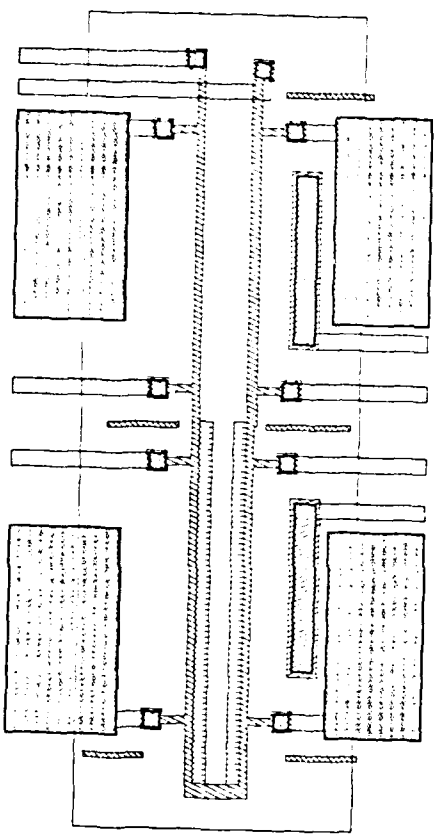


Fig. 4: P+S/D to Active area

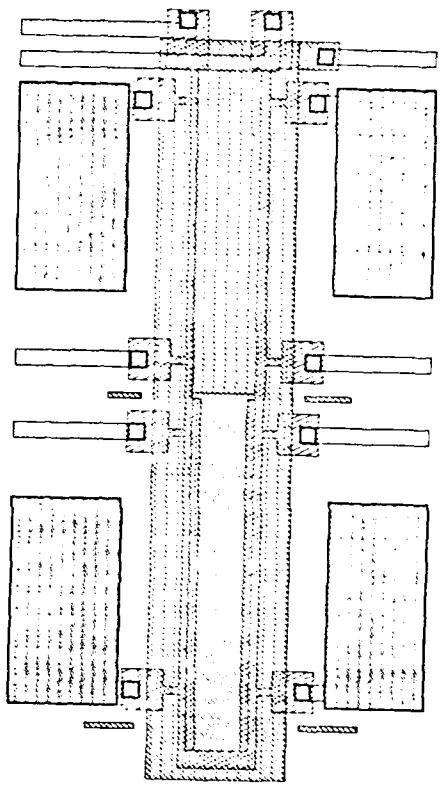


Fig. 5: N-channel implant to Active area

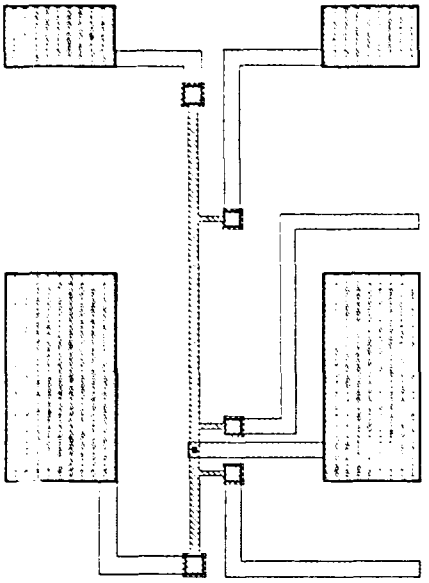


Fig. 6: Contact to Poly

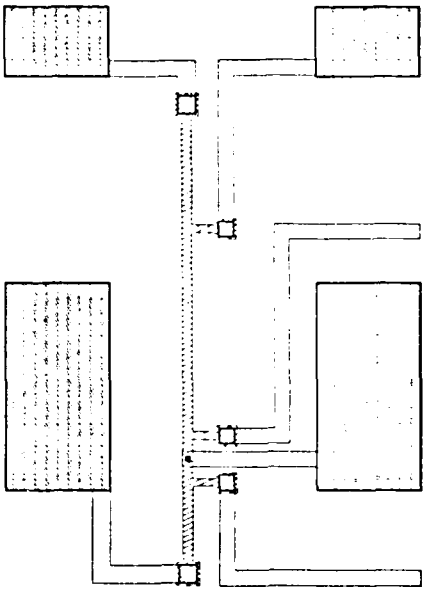


Fig. 7: Contact to Active area

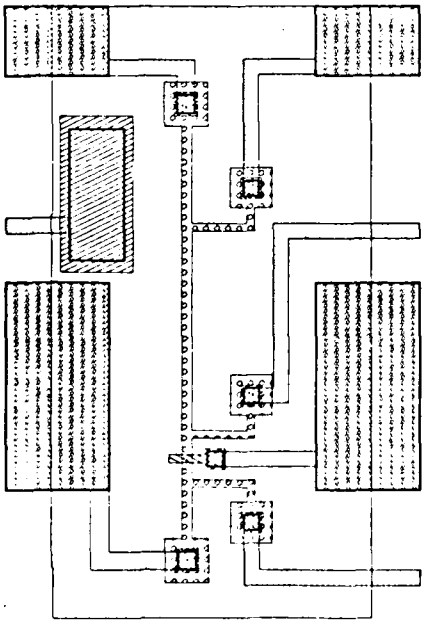


Fig. 8: N-well protect to Active area

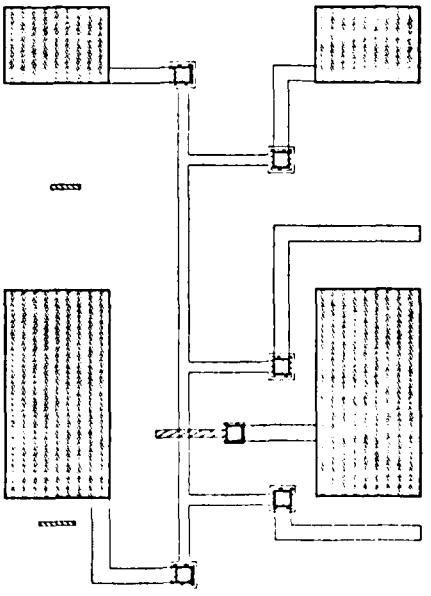


Fig. 9: Active area to N-well

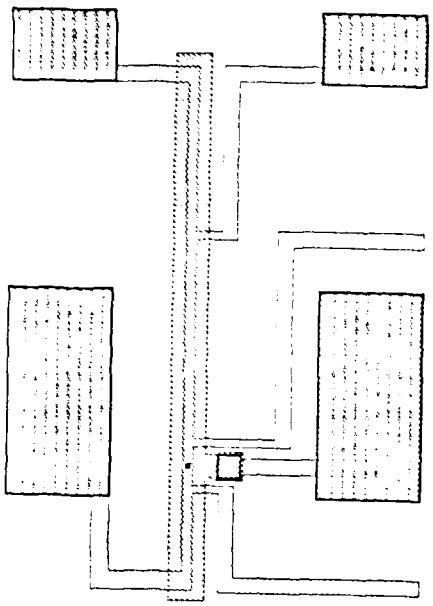


Fig. 10: Metal 1 to Contact

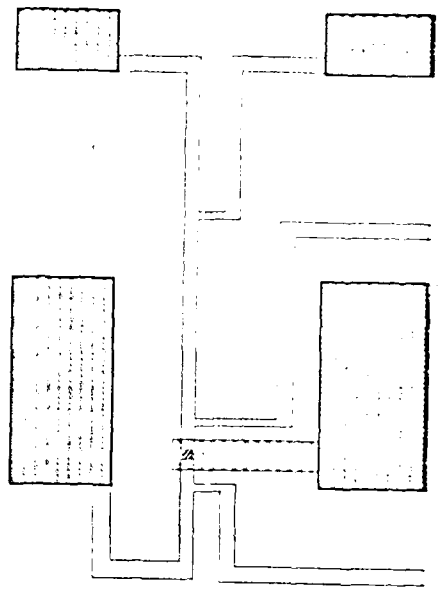


Fig. 11: Via to Metal 1

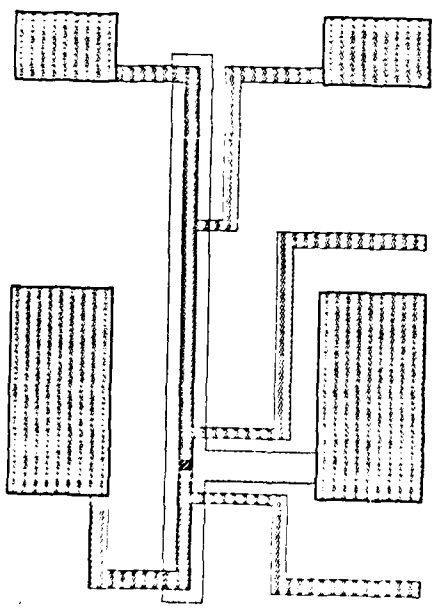


Fig. 12: Metal 2 to Via

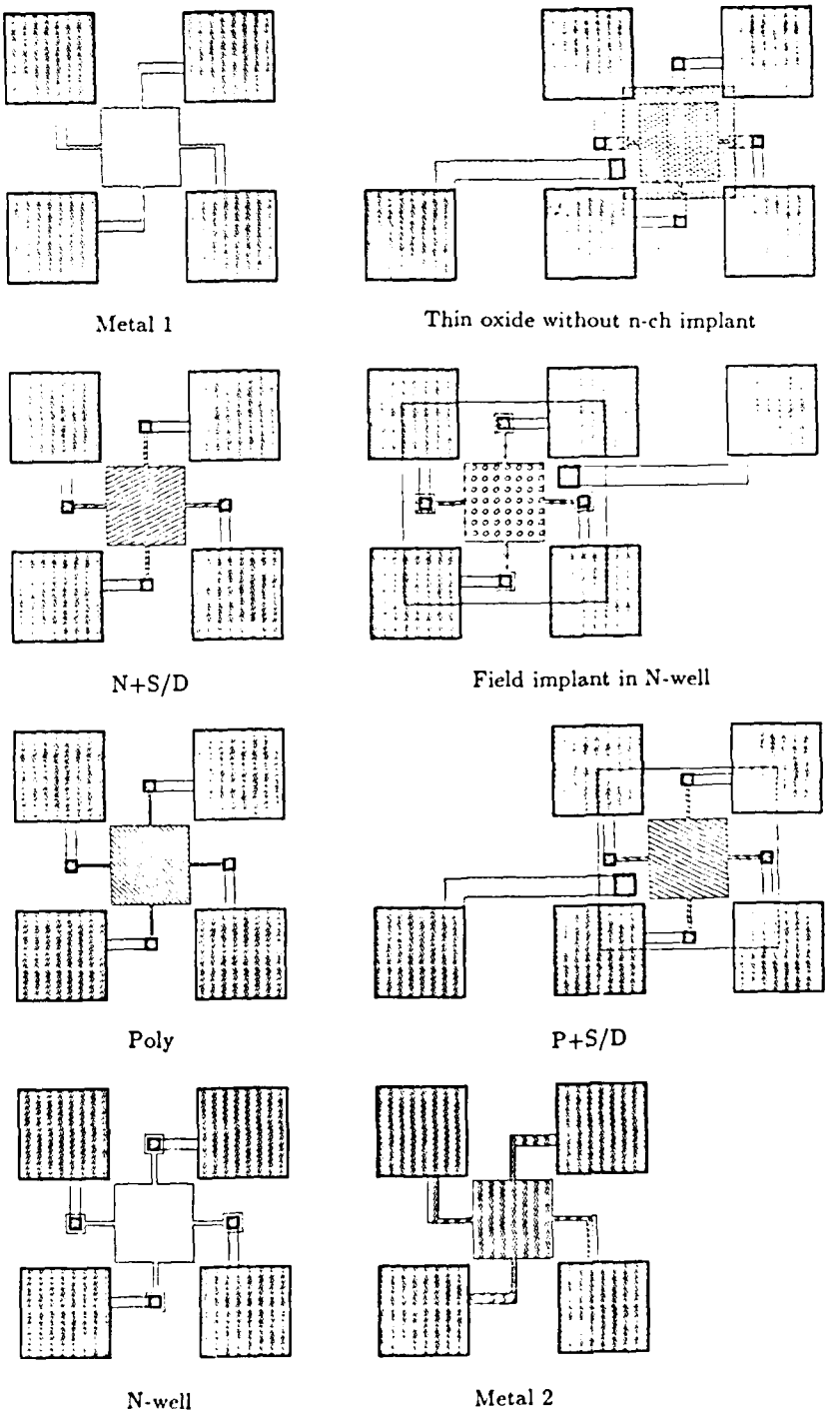


Fig. 13: Van der Pauw sheet resistivity structures

6.4 Measurement and Extraction of Specific Contact Resistivity

W. M. Loh and K. C. Saraswat

6.4.1 Introduction

As MOS and bipolar technologies scale down into the submicron regime, it is believed that the resistances contributed by the contacts increase rapidly and may become an important parameter in determining the performance of the next generation devices. To determine if contact resistance is a limiting factor in the next generation ULSI, it is paramount to obtain accurate values of the specific contact resistivity ρ_c, the physical parameter that governs the interfacial contact resistance between the contact material and the diffusion. There are three test structures commonly used to extract ρ_c: the Cross Bridge Kelvin Resistor (CBKR) [6.4.1], the Contact End Resistor (CER) [6.4.2], and the Transmission Line Tap Resistor (TLTR) [6.4.3]. In operation, a current is sourced from the diffusion level up into the metal level via the contact window. A voltage is measured between the two levels using two other terminals. The contact resistance for each structure is simply this voltage multiplied by the source current. It is important to realize that each device measures the voltage at a different position along the contact, as shown in Fig. 1, hence the resistance values measured are different, and must be clearly defined and distinguished from one another. In this report, they are referred to as R_k (Kelvin), R_e (end), and R_f (front) for the CBKR, CER, and TLTR respectively. From these contact resistance values specific contact resistivity than extracted by using the following relationships:

$$R_k = \frac{V_k}{I} = \frac{\rho_c}{A} \tag{1}$$

for CBKR structure[6.4.1],

$$R_e = \frac{V_e}{I} = R_s \frac{l_t}{W \sinh(l/l_t)} \tag{2}$$

for CER structure[6.4.2] and

$$R_f = \frac{V_f}{I} = R_s \frac{l_t}{W \tanh(l/l_t)} \tag{3}$$

for TLTR structure [6.4.3]. The transfer length l_t is defined as $(\rho_c/R_s)^{1/2}$, a length that characterizes how far the current remains in the diffusion area beneath the contact before passing into the metal level. Eq. (1) - (3) are one dimensional models of the respective test structures for square

contacts of dimension l in a diffusion of width W and sheet resistance R_s. Similar relations have been derived for rectangular contacts and differ only in minor details.

There are two problems which make it very difficult to extract ρ_c accurately from the contact resistance measurements, using the present theoretical 1-D equations The first problem is that when these three structures are used to measure similar contacts, the results often yield conflicting estimates of ρ_c. The second problem is that when contact resistance is plotted against area, a sublinear behavior is observed for CBKR, instead of the expected inverse linear behavior [4-5], as shown in Fig.2 for CBKR structure. The extracted values of ρ_c not only appear to be area dependent, but also a function of diffusion sheet resistance R_s even when the active surface dopant concentrations are the same. This is a serious problem since the variations are often more than an order of magnitude. Recent work [8.5.6-8.5.11] has attributed these phenomena to two-dimensional (2-D) current flow, or crowding, in the diffusion-tap area around the contact window resulting in a parasitic resistance in addition to contact resistance. The value of this resistance has been shown to depend highly on the geometry of the test structure. As a result serious errors can result in the estimation of the values of ρ_c; generally resulting in over estimation of ρ_c. In this report, a unified 2-D model is presented which circumvents these shortcomings and provides an explanation of the two phenomena. The novel extraction method derived from the unified 2-D model provides values of ρ_c which are accurate and self-consistent independent of geometry and test structure types.

6.4.2 Two-Dimensional Model

We shall concentrate here on semiconductor to metal contacts. Since the sheet resistance of metal is much lower than diffusion sheet resistances, metal is considered to be an equipotential plane. Therefore the current flow in the diffusion-contact system can be described entirely by the potential in the diffusion "sheet" which is governed by the Helmholtz equation:

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = \frac{V}{l_t^2} \tag{4}$$

in the diffusion area directly beneath the contact and by the Laplace equation:

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0 \tag{5}$$

elsewhere. For all three test structures, a ratio between a measured potential V^* and the source current I gives a resistance R^*. The resistance ratio R^*/R_s can be expressed as a function of l_t:

$$\frac{R^*}{R}(l_t) = \frac{V^*}{\int_{d_1} \frac{\partial V}{\partial x} dy} \tag{6}$$

where d_1 is a line perpendicular to the current flow. Note that the denominator represents the

total current flowing into the contact window divided by the sheet resistance. Using numerical techniques [6.4.13], we have solved equations (1) and (2) to find $V(x,y,l_t)$ for a wide ranging set of test structure geometries. Then R^*/R_s is evaluated via the solution of (3). By comparing the experimentally measured resistance ratio to the one generated by the 2-D simulations using l_t as a parameter, a unique value of l_t can be extracted and the accurate value of ρ_c obtained for each test structure.

In Fig. 2, the ratio of the Kelvin contact resistance R_k to R_s is shown as a function of contact area for the CBKR structure. The overlap δ, which is the difference between the contact size and diffusion width has been maintained at 2.5 μm and the contact size has been varied. The diagram shows that there is a large deviation from the 1-D model as l_t is decreased. This is due to a parasitic resistance due to current crowding effects added to the ideal value of ρ_c/l^2. This is more pronounced for large contact areas and small values of l_t. For low ρ_c and high R_s, serious crowding occurs and 2-D simulations are required to accurately model the behavior of the data as shown. Good agreement between simulations and measured values implies accurate extraction of different l_t for the metal (W, Al, PtSi) to N^+ and P^+ P, As and B diffusions. The results are summarized in Table 1.

In Fig. 3 the 2-D model is compared to the 1-D model for the contact end resistor for $\delta = 2.5\,\mu$m. The 2-D model shows that the 1-D model severely underestimates this resistance. This is due to current flowing around the overlap area and into the contact end. When the 1-D equation for the CER is used to extract low ρ_c values, the true ρ_c is overestimated. The error can be as high as several orders of magnitude for low values of l_t [6.4.11]. To study the effect of the current flowing in the overlap region, CER test structures were fabricated in which the diffusion-tap width w was varied while the contact size is kept constant at 5 μm. The data and simulations are shown in Fig.4. For curve 1, the value of R_e is independent of w , implying that high value of ρ_c the 1-D model is fairly accurate for. But curves 2 and 3 show a strong dependence on w, which cannot be accounted by the 1-D model according to which ρ_c should be independent of w and should be a function of l^2 only. Once more, the 2-D prediction as shown by the solid lines track the measured data accurately. For PtSi to P diffusion (curve 2), the extracted value of ρ_c is $5 \times 10^{-8}\,\Omega\text{cm}^2$, which agrees well with the values obtained with CBKR and TLTR structures fabricated on the same wafer. But unfortunately, the TLTR structure introduces another source of error, because it can only measure the front resistance indirectly– the sum of front resistance and series diffusion sheet resistances are measured. The "front" potential– the contact potential near the leading edge of the contact in the TLTR structure is larger than the potential on the side as sensed by the CBKR and much larger than the contact end potential in the CER case. Therefore, the 2-D crowding effect has much less impact on the TLTR than the other two structures. In fact, it is possible to design a TLTR which is essentially 1-D by making the contact width much larger than its length. Fig. 5 shows a comparison of 1-D and 2-D models. The 1-D model is off by less than an order of magnitude even for very low values of l_t, signifying that the current crowding effect is not serious for this structure.

The technique described to this point for extracting accurate values of ρ_c is to make several devices with different geometries and one geometric dimension is varied systematically, the R_c/R_s ratios are plotted on the appropriate curve, and ρ_c is extracted by comparing the experimental data with 2-D simulations. The consistency of the extracted ρ_c values should be used as a gauge to judge the validity of the experiment. This technique requires extensive experimental data as well as 2-D simulations for each value of ρ_c. In the next section this constraint is removed by developing

a generalized set of curves.

6.4.3 Generalized Curves For ρ_c Extractions

By using the theory of scaling of contacts [6.4.7] and equations (4) and (5) we have developed a set of generalized curves [6.4.12] . The scaling law allows all of the previous simulations to be normalized, providing three sets of generalized curves as shown in Figures 6-8. In these universal curves the resistance ratio is plotted against the dimensionless ratio l/δ with parameter l_t/δ. These curves eliminate the need for further 2-D simulations since all practical dimensions of contacts and diffusions, and all contact-resistivities and diffusion sheet resistances are contained in one diagram. Fig. 6 shows the case of CBKR structure. When $l_t >> \delta$, the behavior is essentially 1-D and the curves are straight. But as l_t decreases below the 2-D current crowding effect becomes significant, especially in the cases of large contacts. The extraction accuracy in this case is rather low as is shown in the close packing of the curves in this regime of low ρ_c. As contact size shrinks, the curves spread out more evenly and the extraction accuracy is improved. The effect of reducing δ on extraction accuracy is not as obvious because there are two opposing effects. The first being that the l/δ factor will move the operation point to the right and hence decreasing the accuracy whereas the l_t/δ factor will shift the operation point upwards which improves the accuracy. It is the latter which dominates as shown by the close packing and thus for constant contact size, the smaller δ is, the better the extraction accuracy. A similar set of universal curves for CER is shown in Fig. 7. The main difference is that the values of the contact resistance are lower than those of CBKR and this requires difficult measurements of significantly smaller voltages if current densities are kept the same. Furthermore, the resistance ratios decreases rapidly as δ increases and thus the sensitivity of the variation in δ is high. This is because the end contact potential depends almost exponentially on l_t/l . The TLTR is far less sensitive to δ because it detects the front contact potential. Although the solid lines in Fig. 8 indicate that the TLTR has slightly less geometric sensitivities than the CBKR, there exists another source of error caused by the indirect measurement method – extrapolation of the contact resistance from the resistance between two close contacts. When R_s is large, the extrapolation becomes quite difficult and sensitive to small variations in the separation distance between the contacts. But the next example will demonstrate that with careful electrical and optical measurements aided by the universal extraction curves, an accurate value of ρ_c can still be obtained. The R_f/R_s values of the TLTR devices on a single wafer were measured. The separation distances and overlaps were measured by optical and electrical techniques. R_f/R_s data were then plotted as symbols along with the universal curves in Fig. 9. At first glance, the data points appear scattered and poorly fitted to any single curve. But by estimating the value of l_t/δ and then multiplying this by the measured value of δ, these data points all give ρ_c values within 20% of each other. This agreement further strengthens the validity of the generalized method.

6.4.4 Effect of R_s Variation Under the Contact

We have also examined the effect of varying the sheet resistance of the diffusion region below the contact. Contact resistance structures cannot measure specific resistivity directly. What actually is measured is the ratio between contact resistance and sheet resistance. If the sheet resistance below the contact is unknown, then we must instead use a "best guess" to extract resistivity from the data. Usually we can assume that the sheet resistance below the contact is the same as that

measured elsewhere on the die. But there are situations where this may not be true, such as over-etched contacts, or contacts where some semiconductor is consumed during contact formation, e.g., PtSi formation.

Our measurements have not been hampered by the interpretation of this ratio, because we have been more interested in the ratio itself, which is given by the transfer length l_t Changing the sheet resistance below the contact does affect the measurement of l_t however, because the contact current must flow through regions of differing sheet resistance, and the measured voltage is developed by IR drops across both regions.

We have simulated the effect of varying this sheet resistance, and have found the CER to be the most sensitive. On this device, the measured end resistance can increase by as much as 100% when contact sheet resistance is doubled. For the TLTR, the resistance increases by as much as 20%, and for the CBKR, the amount is 10 - 15%. These numbers, of course, depend on the geometry of the structure, but they indicate that the CBKR is the least sensitive to sheet resistance variations.

6.4.5 Comparison of the Three Structures

Using the same design rule for square contacts, the following points are observed. Due to 2-D current crowding, an additional parasitic component is added to the 1-D ideal potential as shown in Fig.9 for each structure. Although the TLTR suffers the least from the parasitics, the front potential V_f can not be obtained directly because it is in the path of the sourcing current. Indirect method such as linear extrapolation to eliminate the diffusion potential drop is required and can result in large errors. The CER, on the other hand, can provide the end potential V_e directly. But unfortunately, the parasitic component usually dominate the potential contribution from the contact end. Furthermore, contrary to the 1-D model, the 2-D current crowding shifts the potential minimum inside the contact as illustrated in Fig.9. These effects make the extraction of ρ_c suffer from difficulties such as extracting the small end potential from much larger parasitic potential and high sensitivities to geometrical parameters such as diffusion width and overlap δ. Judging from the arguments presented above, the CBKR emerges as the best compromise between extraction ease and sensitivities when aided with 2-D simulations. The Kelvin potential is less accurate compared to the front potential but superior to the end potential with the advantage that it can be measured directly. In conclusion, we recommend the CBKR above the TLTR and CER in the extraction of ρ_c for the next generation of ULSI.

6.4.6 ρ_c vs. Surface Dopant Concentration

A large number of test devices have been fabricated with PtSi, Pd_2Si and selective CVD W as the contact materials. Surface dopant concentration for N^+ and P^+ diffusions was varied over a wide range. ρ_c was extracted from the V/I measurements done on carefully designed test structures. Figure 10 shows the results for the case of CVD tungsten. These values are independent of test structure type and contact geometry and size. It can be seen that ρ_c values in the range of $10^{-8}\,\Omega\text{-cm}^2$ can be obtained for both type of diffusions. As a result it appears that even for submicron MOS devices the contact resistance should not limit the performance.

6.4.7 Conclusions

We have shown an extraction technique which allows accurate determination of contact resistivity even at low values, and which repeatedly returns the same value of ρ_c for a wide variety of contact dimensions and test structures on a single wafer. The results of the 2-D model are presented in a universal form, so that they may be used for extractions without performing any additional simulations. The new model points out that the 1-D model seriously overestimates the specific contact resistance due to 2-D current crowding in the overlap region. This implies that most reported values of ρ_c extracted by using the 1-D method are overestimated, and that ρ_c will not be a limiting factor for ULSI.

6.4.8 References

[6.4.1] . S. J. Proctor , L. W. Linholm and J. A. Mazer , " Direct Measurement of Interfacial Contact Resistance , End Contact Resistance, and Interfacial Contact Layer Uniformity, " IEEE Trans. Electron Devices, vol. ED-30 , p.1535-1542 , 1983.

[6.4.2] . J. Chern and W. G. Oldham , " Determining Specific Contact Resistivity from Contact End Resistance Measurements , " IEEE Electron Device Lett. , vol. EDL-5 , p.178-180 , May 1984.

[6.4.3] . H. H. Berger, " Contact Resistance and Contact Resistivity , " Electrochem. Soc. Journal, vol.119, no. 4, p.507-514, Apr. 1972.

[6.4.4] . R. L. Maddox, " On the Optimization of VLSI Contacts, " IEEE Trans. Electron Devices, vol. ED-32, no. 3, p.682-690, Mar. 1985.

[6.4.5] . J. M. Ford, "AlSiContact Resistance for Submicrometer Design Rules, " IEEE Trans. Electron Devices, vol. ED-32, no. 4, p.840-842, Apr. 1985.

[6.4.6] . M. Finetti, A. Scorzoni, and G. Soncini, " Lateral Current Crowding Effects on Contact Resistance Measurements in Four Terminal Resistor Test Patterns, " IEEE Electron Device Lett. , vol. EDL-5, no. 12, p.524-526, Dec. 1984.

[6.4.7] . W. M. Loh, K. Saraswat, and R. W. Dutton, " Analysis and Scaling of Kelvin Resistors for Extraction of Specific Contact Resistivity, " IEEE Electron Device Lett., vol. EDL-6, no. 3, p.105-108, Mar. 1985.

[6.4.8] . M. Finetti, A. Scorzoni, and G. Soncini, " A Further Comment on ' Determining Specific Contact Resistivity from Contact End Resistance Measurements ', " IEEE Electron Device Lett., vol. EDL-6, no. 4, p.184-185, Apr. 1985.

[6.4.9] . N. P. Armstrong, and K. R. Stribley, " Development of Low-Resistive Contacts in VLSI Using 2-D Simulation, " 1985 Proceedings of Second International IEEE VLSI Multilevel Interconnection Conference, Santa Clara, p. 389-395, June 1985.

[6.4.10] . W. M. Loh, S. E. Swirhun, E. Crabbe, K. Saraswat, and R. M. Swanson, " An accurate Method to Extract Specific Contact Resistivity Using Cross Bridge Kelvin Resistors, " IEEE Electron Device Lett., vol. EDL-6, no. 9, p.441-443, Sept. 1985.

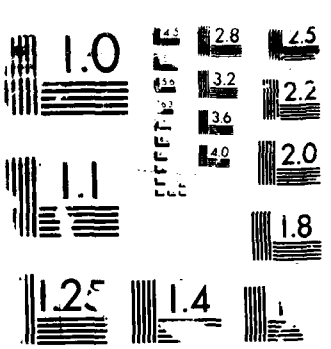
[6.4.11] . S. E. Swirhun, W. M. Loh, R. M. Swanson, and K. Saraswat, " Current Crowding Effects and Determination of Specific Contact Resistivity from Contact End Resistance (CER) Measurement" , IEEE Elect. Dev. Letters, vol. EDL-6, p.639-641, Dec. 1985.

[6.4.12] .
.W. M. Loh, S. E. Swirhun, T. A. Schreyer, R. M. Swanson, and K. C. Saraswat, "2-D Simulations for Accurate Extraction of the Specific Contact Resistivity from Contact Resistance Data" Technical Digest of 1985 IEEE IEDM, pp.586-589.

[6.4.13] . M. Pinto, C. Rafferty, R. W. Dutton, " PISCES-II: Poisson and Continuity Equation Solver, " Stanford Electronics Laboratories Technical Report, October 1984.

AD-A189 451 COMPUTER AIDED FAST TURNAROUND LABORATORY FOR RESEARCH IN VLSI (VERY LARG (U) STANFORD UNIV CA CENTER FOR INTEGRATED SYSTEMS J D MEINDL ET AL 31 MAY 87 1/1

UNCLASSIFIED MDA903-84-K-0062 F/G 9/1 NL


1.0
1.1
1.25
1.4
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

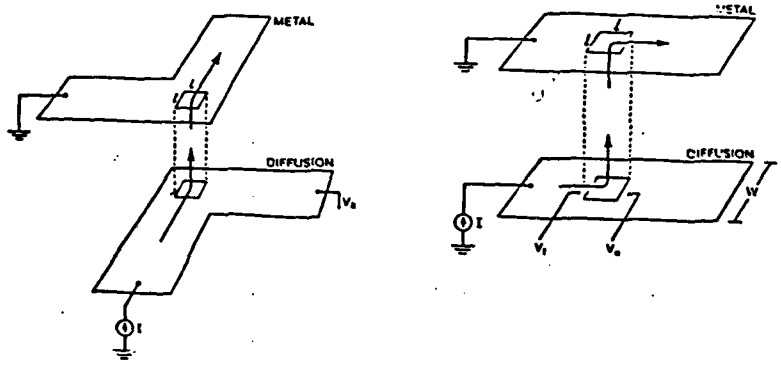


Fig.1 Basic principle of the contact resistance measurement using (a) cross bridge Kelvin resistor (b) Contact end resistor transmission line tap resistor.

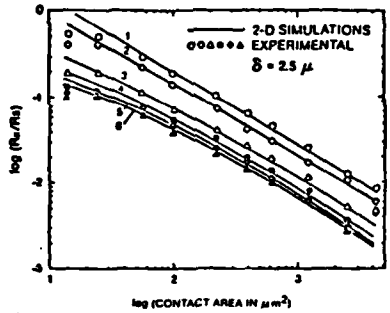


Fig. 2 The ratio of Kelvin contact resistance to diffusion sheet resistance plotted as a function of contact area for different transfer length l, for CBKR with overlap being 2.5 μm. Symbols denote measured data, solid lines are simulations.

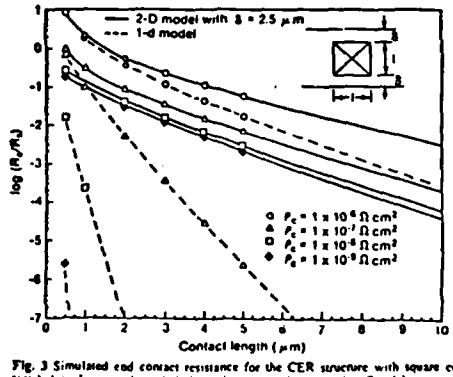


Fig. 3 Simulated end contact resistance for the CER structure with square contact holes. Lines and symbols both denote simulations only. Solid lines are the 2-D model; dashed lines are the 1-D model.

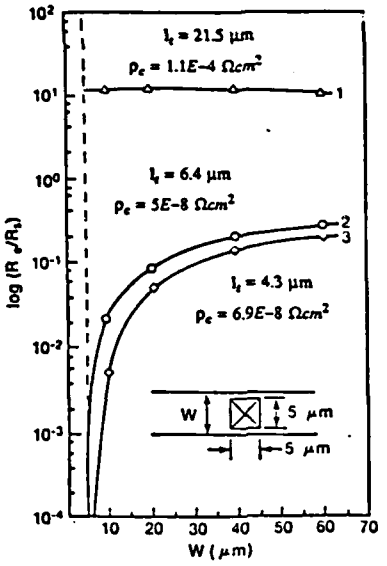


Fig. 4 End resistance as a function of diffusion width for constant contact size of 5 μm. Symbols denote actual measurements; lines are 2-D simulations.

Table 1 Summary of extraction results for the data shown in Fig. 2.

Num.	Metal	Dopant	R_s (Ω/sq.)	l_t (μm)	ρ_c (Ωμm²)
1	W	As	422	3.5	3500
2	W	B	71.4	2.75	540
3	Al	B	12.4	1.4	24.3
4	Al	B	71.1	0.9	57.6
5	PtSi	P	12.1	0.65	5.11
6	W	As	44.4	0.5	11.1

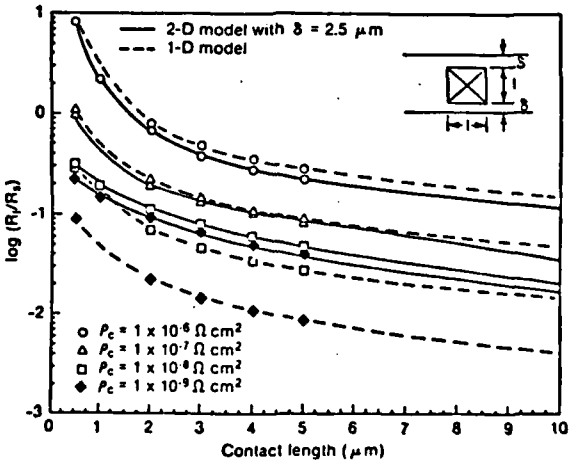


Fig. 5 Simulated front contact resistance for the TLTR structure with square contact holes. Lines and symbols both denote simulations only. Solid lines are the 2-D model; dashed lines are the 1-D model.

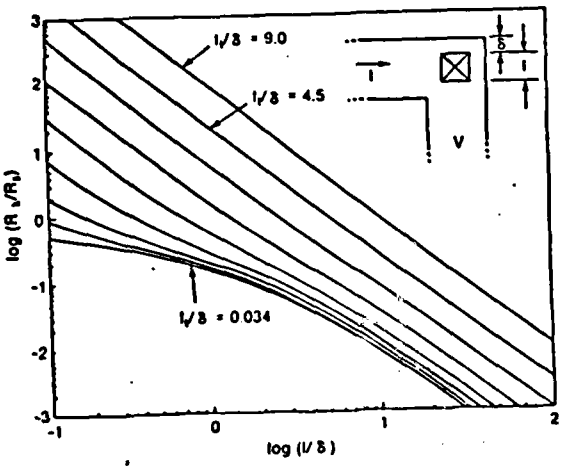


Fig. 6 Generalized universal curves for the CBKR. Curves show l_i / δ in octave steps.

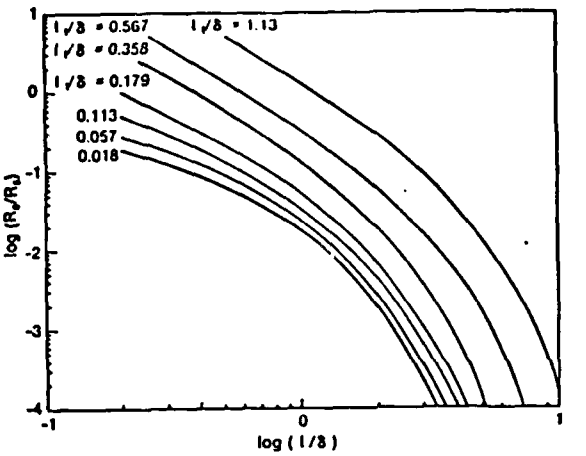


Fig. 7 Generalized universal curves for the CER. Curves show l_i / δ approximately in octave steps.

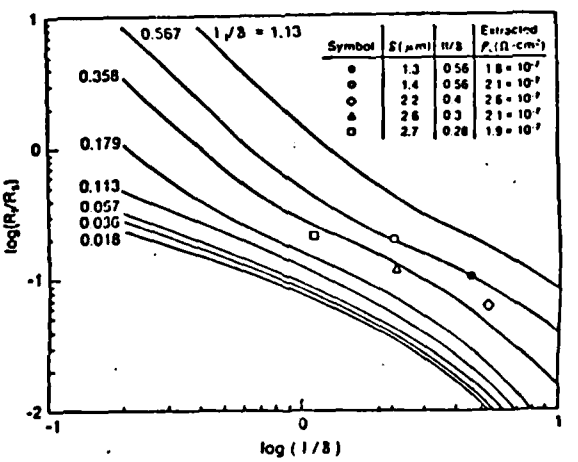


Fig. 8 Generalized universal curves for the TLTR. Curves show l_i / δ in approximate octave steps. Symbols show data from devices on the same wafer which have a variety of overlap sizes.

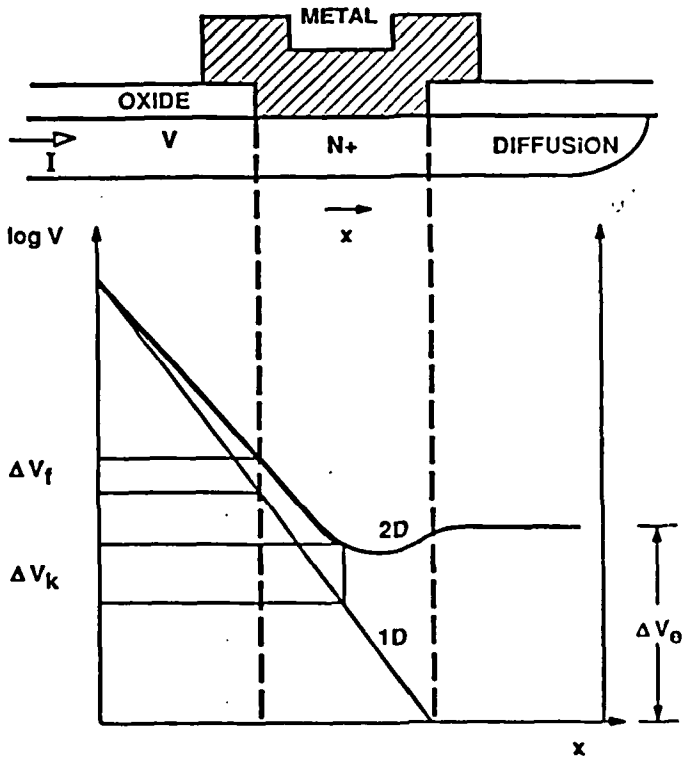


Fig. 9 A qualitative illustration of the potential distribution of the contact system.

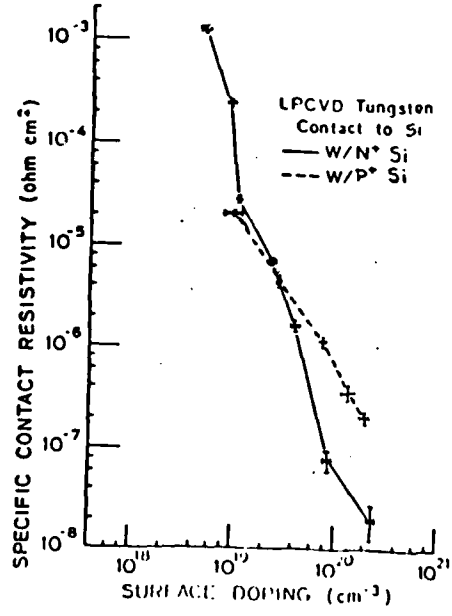


Fig. 10 Specific contact resistivity of LPCVD tungsten to silicon contacts

6.5. Ontology of CMOS SPICE Parameters

6.5.1. Introduction

VLSI fabrication is a very long and complex process which involves more than one hundred process steps and requires more than ten masks for layer definitions. In this lengthy process, small fluctuations in various steps can aggregate to cause the performance of the final circuit to deviate from the design specifications. Stabilizing these circuit parameters and closely maintaining them is a crucial problem in VLSI manufacturing.

A system which will monitor and help correct deviations in the process could be developed if we could infer from parametric measurements what changes in the process need to be made. To this end, Pan [5] divided the parameters involved in parametric test data interpretation into 3 levels:

1. **Effective Level:** also called the *measurement level*, consists of the electrically measurable parameters.
2. **Self Level:** also called *physical structure level*, consists of the physical parameters like junction depth, substrate doping concentration, oxide thickness, etc.
3. **Cause Level:** also called *process level*, consists of the process parameters like oxidation time, diffusion time and temperature, etc.

A knowledge base called PIES was built to make inferences on the 3 levels of parameters mentioned above. The goal of PIES is to diagnose process problems by analyzing parametric test data. From measured electrical parameters, PIES infers the quality of physical parameters, and then it deduces the quality of process parameters (e.g. oxidation time too long, diffusion temperature too high, etc). Basically, PIES forms a frame for the knowledge about IC production line diagnosis.

The relation between physical parameters and process parameters has also been studied. Mohammed [4] describes the knowledge and representation needed for an expert system that can qualitatively simulate the effect of the fabrication process on device physical parameters. His study sets up the relations between levels 2 and 3 above. A method to bridge between levels 1 and 2 remains to be built, and this is the main effort of this research.

As was previously mentioned, the parameters that should be monitored and maintained are the circuit simulation parameters, such as those used in SPICE. If we can maintain the SPICE parameters, we can stabilize the performance of the VLSI chips. Therefore, we wish to start with a list of the SPICE parameters as the level 1 measured parameters and develop relations between them and the level 2, or physical structure of the devices. This represents an "ontology study" which is simply an ordering of knowledge about SPICE models. Once we relate the SPICE parameters with the level 2 parameters, we can reach the level 3 parameters through the bridge established by Mohammed. Then we would have a complete linkage between SPICE parameters and process parameters.

In this paper, we first describe the newest CMOS model used in the SPICE circuit simulation

program. Second, we briefly describe the method used to represent the knowledge obtained from the this model. Then the knowledge about CMOS SPICE parameters is put together as a mini knowledge base. We also give some examples of how to make inferences from the knowledge base. Last, we propose a view of the prospective system configuration and the work that can be done in the future.

6.5.2. BSIM Model for MOS transistors

The newest Berkeley SPICE MOS model is called BSIM. As input, it requires parameters concerning small geometry effects like depletion charge sharing, mobility degradation, velocity saturation, drain induced barrier lowering, etc. Table 6.5.1 lists the parameters needed in the BSIM model for SPICE to do level 4 simulations. [8].

A method to obtain the value of the BSIM parameters from an IC production line is described in [7]. The method is simple: the parameters are measured on transistors of different geometry and curve fitting is used to obtain geometry independent parameters. These geometry independent parameters depend only on the process. They are put into a "process file", which can be combined with a device size specification file to compute the parameters needed for a SPICE simulation. The procedure is shown in Figure 6.5.1.

In reality, the "process file" parameters are not fixed. They vary across the wafer, between wafers , and between process runs and form statistical distributions. Our goal is to help make the parameter distributions as narrow as possible. In addition, we wish to help control the production line when any of the parameters deviates from its optimal value.

In order to maintain the SPICE parameters (in the "process file") through process control, it will be necessary to understand the relationship between each SPICE parameter and the underlying device physical or structural parameters. The first step in determining this relationship is to list each SPICE BSIM parameter and obtain a quantitative relationship between it and the physical parameters. The following equations were compiled from several common textbooks on integrated circuits [2] [3].

$\sim$ *means proportional to.*

$\simeq$ *means approximately equal.*

name	parameter	units	l/w
VFB	flat-band voltage	V	*
PHI	surface inversion potential	V	*
K1	body effect coefficient	$V^{1/2}$	*
K2	drain/source depletion charge sharing coefficient	-	*
ETA	zero-bias drain-induced barrier lowering coefficient	-	*
MUZ	zero-bias mobility	$cm^2/V-s$	
DL	shortening of channel	µm	
DW	narrowing of channel	µm	
U0	zero-bias transverse-field mobility degradation coefficient	V^{-1}	*
U1	zero-bias velocity saturation coefficient	µm/V	*
X2MZ	sens. of mobility to substrate bias at v_{ds}=0	cm^2/V^2-s	*
X2E	sens. of drain-induced barrier lowering effect to substrate bias	V^{-1}	*
X3E	sens. of drain-induced barrier lowering effect to drain bias at $V_{ds}=V_{dd}$	V^{-1}	*
X2U0	sens. of transverse field mobility degradation effect to substrate bias	V^{-2}	*
X2U1	sens. of velocity saturation effect to substrate bias	μmV^{-2}	*
MUS	mobility at zero substrate bias and at $V_{ds}=V_{dd}$	cm^2/V^2-s	
X2MS	sens. of mobility to substrate bias at $V_{ds}=V_{dd}$	cm^2/V^2-s	*
X3MS	sens. of mobility to drain bias at $V_{ds}=V_{dd}$	cm^2/V^2-s	*
X3U1	sens. of velocity saturation effect on drain bias at $V_{ds}=V_{dd}$	μmV^{-2}	*
TOX	gate oxide thickness	µm	
TEMP	temperature at which parameters were measured	°C	
VDD	measurement bias range	V	
CGDO	gate-drain overlap capacitance per meter channel width	F/m	
CGSO	gate-source overlap capacitance per meter channel width	F/m	
CGBO	gate-bulk overlap capacitance per meter channel length	F/m	
XPART	gate-oxide capacitance charge model flag	-	
N0	zero-bias subthreshold slope coefficient	-	*
NB	sens. of subthreshold slope to substrate bias	-	*
ND	sens. of subthreshold slope to drain bias	-	*
RSH	drain and source diffusion sheet resistance	Ω/□	
JS	source drain junction current density	A/m^2	
PB	built in potential of source drain junction	V	
MJ	Grading coefficient of source drain junction	-	
PBSW	built in potential of source,drain juntion sidewall	V	
MJSW	grading coefficient of source drain junction sidewall	-	
CJ	Source drain junction capacitance per unit area	F/m^2	
CJSW	source drain junction sidewall capacitance per unit length	F/m	
WDF	source drain junction default width	m	
DELL	Source drain junction length reduction	m	

Table 6.5.1: SPICE BSIM Parameters

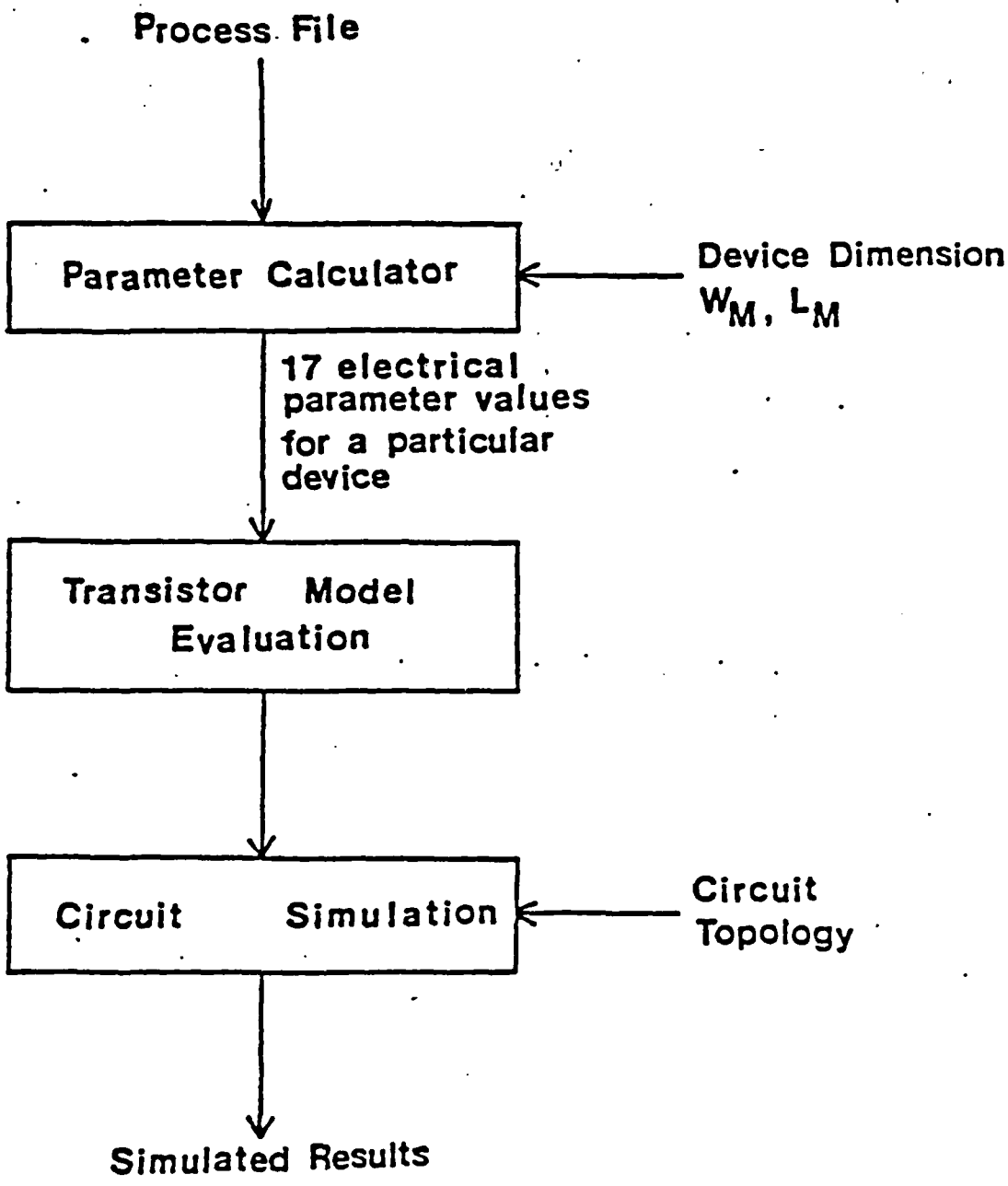


Figure 6.5.1: Parameter Manipulations in a Circuit Simulator

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = k' \frac{W_{eff}}{L_{eff}} (V_{GS} - V_{th})(1 + \lambda V_{DS}) \tag{6.5.1}$$

$$k' = \mu C_{ox} \tag{6.5.2}$$

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{6.5.3}$$

$$\lambda = \frac{1}{V_A} = \frac{\partial X_d / \partial V_{DS}}{L_{eff}} = \frac{X_d}{2L_{eff}(V_{DS} - V_{DS(sat)})} \tag{6.5.4}$$

$$X_d \sim \sqrt{2\epsilon_s (V_{DS} - V_{DS(sat)}) / qN_B} \tag{6.5.5}$$

$$g_{mb} = g_m \frac{\gamma}{2\sqrt{\phi_s - V_{BS}}} \tag{6.5.6}$$

$$\gamma = \frac{\sqrt{2qN_B\epsilon_s}}{C_{ox}} \tag{6.5.7}$$

$$r_o = (\frac{\partial I_D}{\partial V_{DS}})^{-1} = \frac{1}{\lambda I_D} \tag{6.5.8}$$

$$C_{sb/db} = \frac{C_{sb0/db0}}{(1 + V_{SB/DB} / \psi_0)^{1/2}} \tag{6.5.9}$$

$$C_{sb0/db0} = A[\frac{q\epsilon_s N_B N_{S/D}}{2(N_B + N_{S/D})}]^{1/2} \frac{1}{\psi_0} \tag{6.5.10}$$

$$\psi_0 = kT \ln \frac{N_{S/D} N_B}{n_i^2} \tag{6.5.11}$$

$$C_{gs} \simeq W_{eff} C_{gs0} + \frac{2}{3} W_{eff} L_{eff} C_{ox} \tag{6.5.12}$$

$$W_{eff} = W - 2\Delta W \tag{6.5.13}$$

$$L_{eff} = L - 2\Delta L \tag{6.5.14}$$

$$C_{gd} \simeq W_{eff} C_{gd0} \tag{6.5.15}$$

$$C_{gs0/gd0} \sim C_{ox} \Delta L \tag{6.5.16}$$

$$C_{gb} \sim C_{fox} \Delta W \tag{6.5.17}$$

$$C_{fox} = \frac{\epsilon_{ox}}{t_{fox}} \tag{6.5.18}$$

$$R_{sh} = \frac{1}{\mu N_S D_j} \tag{6.5.19}$$

All the equations listed above can be applied to the BSIM model except (6.5.1), which must be modified. As described in [7], g_m takes the following form.

$$g_m = \frac{\beta(V_{GS} - V_{th})}{2aK} \tag{6.5.20}$$

where a is the conductance-degradation coefficient, and K is a factor used to transform (6.5.1) to (6.5.20) [7].

$$\beta = \frac{\beta_0}{1 + U_0(V_{GS} - V_{th})} \tag{6.5.21}$$

where U_0 is mobility degradation coefficient.

$$\beta_0 = \mu C_{ox} W_{eff} / L_{eff} \tag{6.5.22}$$

$$a = 1 + \frac{gK_1}{2\sqrt{\phi_s - V_{BS}}} \tag{6.5.23}$$

where g is a fitting factor, and K_1 is the body effect coefficient.

$$g = 1 - \frac{1}{1.744 + 0.8364(\phi_s - V_{BS})} \tag{6.5.24}$$

$$K_1 = \frac{\sqrt{2q\epsilon_s N_B}}{C_{ox}} \tag{6.5.25}$$

$$K = \frac{1 + v_c + \sqrt{1 + 2v_c}}{2} \tag{6.5.26}$$

$$v_c = \frac{U_1(V_{GS} - V_{th})}{a} \tag{6.5.27}$$

$$U_1 = \frac{1}{E_{crit} L_{eff}} \tag{6.5.28}$$

$$V_{th} = V_{FB} + \phi_s + K_1\sqrt{\phi_s - V_{BS}} - K_2(\phi_s - V_{BS}) - \eta V_{DS} \tag{6.5.29}$$

where V_{FB} is the flat band voltage, ϕ_s is the surface potential when an inversion layer is formed, K_1 is the body effect coefficient, K_2 is the charge sharing coefficient, and η is the drain induced barrier lowering coefficient.

$$V_{FB} = \phi_{ms} - \frac{qN_{ss}}{C_{ox}} - \frac{1}{C_{ox}}\int_0^{t_{ox}} \frac{x}{t_{ox}} \rho(x) dx \tag{6.5.30}$$

where ϕ_{ms} is a function of poly doping and substrate doping concentration, N_{ss} is the interface charge density, and ρ is a function of charge distribution in the oxide.

$$\phi_s = 2\frac{kT}{q} ln(\frac{N_B}{n_i}) \tag{6.5.31}$$

Ratnakumar [6] derived expressions for K_2 and η.

$$K_2 = \frac{12t_{ox}}{d_1} e^{-(\pi L_{eff}/4d_1)} \tag{6.5.32}$$

$$\eta = \frac{6t_{ox}}{d_1} e^{-(\pi L_{eff}/4d_1)} \tag{6.5.33}$$

where d_1 is the depth of the threshold adjustment implant, which is a function of surface

concentration N_{suf} and bulk concentration N_B.

6.5.3. Representation of Knowledge

Two different levels of knowledge can be used in making the transition from the measurement level to the physical level for the purpose of process debugging.

1. **Quantitative Relations:** The equations above can be considered as a set of simultaneous equations. After collecting the necessary amount of data, the system can solve the above equations for the physical parameters and deduce which ones have changed in order to find the correction at the cause level. This, however, will be a very cumbersome if not impossible task considering the complexity of the equations. In addition, all of the above equations represent only approximations to real life. Therefore, even such a rigorous approach is subjected to uncertainties.

2. **Qualitative Relations:** For each measured parameter, we can determine simple qualitative dependencies on the physical parameters and other measured parameters from the above equations. This information can be put into a knowledge base for qualitative reasoning. While this may not be as accurate as a method of qualitative relations, it is a much more manageable task for the system.

   A combined approach suggests that we can do the qualitative reasoning first to find the candidates of the solution, then do quantitative calculations on those candidates. The candidates generated from qualitative reasoning might be a very small fraction of all possible parameters, and hence the load for quantitative calculation would be greatly decreased.

Figure 6.5.2 shows the dependencies among parameters. Note in this figure, *y(x1,x2,...)* means *y* depends on *x1 ,x2...*, where *y, x1, x2...* are parameters of a transistor.

In addition, the parameters are different for different types of transistors. To distinguish between PMOS and NMOS parameters, we can write down the following relations:

CMOS(PMOS, NMOS)
PMOS(g_m, g_{mb}, r_o, C_{gd}, C_{gs}, C_{gb}, C_{sb}, C_{db}, R_{sh})
NMOS(g_m, g_{mb}, r_o, C_{gd}, C_{gs}, C_{gb}, C_{sb}, C_{db}, R_{sh}).

Many knowledge representation methods exists, of which predicate calculus is the most suitable for rule based systems.

$g_m(\beta, a, K, V_{th})$
- $\beta(\beta_0, U_0, V_{th})$
  - $\beta_0(\mu, C_{ox}, W_{eff}, L_{eff})$
    - $C_{ox}(t_{ox})$
    - $W_{eff}(W, \Delta W)$
    - $L_{eff}(L, \Delta L)$
- $a(K_1, \phi_s)$
- $K(U_1, a, V_{th})$
- $V_{th}(V_{FB}, \phi_s, K_1, K_2, \eta)$
  - $V_{FB}(C_{ox}, \phi_{ms}, N_{ss}, Na)$
  - $\phi_s(N_B)$
  - $K_1(N_B, C_{ox})$
  - $K_2(t_{ox}, d_1, L_{eff})$
    - $d_{eff}(N_{suf}, N_B)$
  - $\eta(t_{ox}, d_1, L_{eff})$

$g_{mb}(g_m, \gamma)$
- $\gamma(N_B, C_{ox})$

$r_o(\lambda)$
- $\lambda(X_d, L_{eff})$
  - $X_d(N_B)$

$C_{gs}(C_{gs0}, C_{ox}, W_{eff}, L_{eff})$
- $C_{gs0}(C_{ox}, \Delta L)$

$C_{gd}(C_{gs0}, W_{eff})$
- $C_{gd0}(C_{ox}, \Delta L)$

$C_{gb}(C_{fox}, \Delta W)$
- $C_{fox}(t_{fox})$

$C_{sb}(C_{sb0}, \psi_0)$
- $C_{sb0}(N_S, N_B, \psi_0)$
- $\psi_0(N_S, N_B)$

$C_{db}(C_{db0}, \psi_0)$
- $C_{db0}(N_S, N_B, \psi_0)$

$R_{sh}(\mu, N_S, D_j)$

Figure 6.5.2: Relationships Between Parameters

6.5.4. A Mini Knowledge Base for CMOS SPICE Parameters

In this section, we put the relationships between parameters listed in section 3 into predicate calculus rules, and form a mini knowledge base for CMOS SPICE parameters.

The objects and relations in our conceptualization of the knowledge about CMOS SPICE parameters are listed below.

Types of objects:

- CMOS devices
- PMOS transistors
- NMOS transistors
- Transistor parameters: g_m, g_{mb}, V_{th}, μ, etc.
- Dependency qualities: Co_variant, Contra_variant, Null.

Types of relations:

- Cmos(t): means t is a CMOS device.
- Pmos(t): means t is a PMOS transistor.
- Nmos(t): means t is an NMOS transistor.
- α_of(x,t): means x is the α of t, eg. Gm_of(x,t) means x is the g_m of t; Gmb_of(x,t) means x is the g_{mb} of t, etc.
- Vary_with(x,y,z): means x varies with y of type z. z could be Co_variant (means x and y vary in the same direction), Contra_variant (means x and y vary in the opposite direction), or Null (means unclear or not applicable).

With the objects and relations defined above, we can code the knowledge about CMOS SPICE parameters in predicate calculus rules. They are shown in Figure 6.5.3. These rules may easily be transformed into a Prolog or Lisp program.

P1: $\forall\, t1,t2,t3\ Cmos(t1) \wedge Pmos\_of(t2,t1) \wedge Nmos\_of(t3,t1)$
$\rightarrow Vary\_with(t1,t2,Null) \vee Vary\_with(t1,t3,Null)$

P2: $\forall\, t,x1,x2,x3,x4,x5,x6,x7,x8,x9\ Pmos(t) \wedge Gm\_of(x1,t)$
$\wedge Gmb\_of(x2,t) \wedge Ro\_of(x3,t) \wedge Cgd\_of(x4,t)$
$\wedge Cgs\_of(x5,t) \wedge Cgb\_of(x6,t) \wedge Csb\_of(x7,t)$
$\wedge Cdb\_of(x8,t) \wedge Rsh\_of(x9,t)$
$\rightarrow Vary\_with(t,x1,Null) \vee Vary\_with(t,x2,Null)$
$\vee Vary\_with(t,x3,Null) \vee Vary\_with(t,x4,Null)$
$\vee Vary\_with(t,x5,Null) \vee Vary\_with(t,x6,Null)$
$\vee Vary\_with(t,x7,Null) \vee Vary\_with(t,x8,Null)$
$\vee Vary\_with(t,x9,Null)$

P3: $\forall\, t,x1,x2,x3,x4,x5,x6,x7,x8,x9\ Nmos(t) \wedge Gm\_of(x1,t)$
$\wedge Gmb\_of(x2,t) \wedge Ro\_of(x3,t) \wedge Cgd\_of(x4,t)$
$\wedge Cgs\_of(x5,t) \wedge Cgb\_of(x6,t) \wedge Csb\_of(x7,t)$
$\wedge Cdb\_of(x8,t) \wedge Rsh\_of(x9,t)$
$\rightarrow Vary\_with(t,x1,Null) \vee Vary\_with(t,x2,Null)$
$\vee Vary\_with(t,x3,Null) \vee Vary\_with(t,x4,Null)$
$\vee Vary\_with(t,x5,Null) \vee Vary\_with(t,x6,Null)$
$\vee Vary\_with(t,x7,Null) \vee Vary\_with(t,x8,Null)$
$\vee Vary\_with(t,x9,Null)$

P4: $\forall\, t,x1,x2,x3,x4,x5\ Gm\_of(x1,t) \wedge Beta\_of(x2,t) \wedge A\_of(x3,t)$
$\wedge K\_of(x4,t) \wedge Vth\_of(x5,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Contra\_variant)$
$\vee Vary\_with(x1,x4,Contra\_variant) \vee Vary\_with(x1,x5,Null)$

P5: $\forall\, t,x1,x2,x3,x4\ Beta\_of(x1,t) \wedge Beta0\_of(x2,t) \wedge U0\_of(x3,t)$
$\wedge Vth\_of(x4,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Contra\_variant)$
$\vee Vary\_with(x1,x4,Null)$

P6: $\forall\, t,x1,x2,x3,x4,x5\ Beta0\_os(x1,t) \wedge Mu\_of(x2,t) \wedge Cox(x3,t)$
$\wedge Weff\_of(x4,t) \wedge Leff\_of(x5,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$
$\vee Vary\_with(x1,x4,Co\_variant) \vee Vary\_with(x1,x5,Contra\_variant)$

P7: $\forall\, t,x1,x2\ Cox\_of(x1,t) \wedge Tox\_of(x2,t)$
$\rightarrow Vary\_with(x1,x2,Contra\_variant)$

Figure 6.5.3 : Predicate Calculus Representations of Knowledge about CMOS SPICE Parameters

P8: $\forall t,x1,x2,x3\ Weff\_of(x1,t) \wedge W\_of(x2,t) \wedge DW\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Contra\_variant)$

P9: $\forall t,x1,x2,x3\ Leff\_of(x1,t) \wedge L\_of(x2,t) \wedge DL\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Contra\_variant)$

P10: $\forall t,x1,x2,x3\ A\_of(x1,t) \wedge K1\_of(x2,t) \wedge PhiS\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Contra\_variant)$

P11: $\forall t,x1,x2,x3,x4\ K\_of(x1,t) \wedge U1\_of(x2,t) \wedge A\_of(x3,t) \wedge Vth\_of(x4,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Contra\_variant)$
$\vee Vary\_with(x1,x4,Null)$

P12: $\forall t,x1,x2,x3,x4,x5,x6\ Vth\_of(x1,t) \wedge Vfb\_of(x2,t) \wedge PhiS\_of(x3,t)$
$\wedge K1\_of(x4,t) \wedge K2\_of(x5,t) \wedge Eta\_of(x6,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$
$\vee Vary\_with(x1,x4,Co\_variant) \vee Vary\_with(x1,x5,Contra\_variant)$
$\vee Vary\_with(x1,x6,Contra\_variant)$

P13: $\forall t,x1,x2,x3,x4,x5\ Vfb\_of(x1,t) \wedge Cox\_of(x2,t) \wedge PhiMs\_of(x3,t)$
$\wedge Nss\_of(x4,t) \wedge Na\_of(x5,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$
$\vee Vary\_with(x1,x4,Contra\_variant)$
$\vee Vary\_with(x1,x5,Contra\_variant)$

P14: $\forall t,x1,x2,x3\ PhiMs\_of(x1,t) \wedge Npoly\_of(x2,t) \wedge [illegible]\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,,Null) \vee Vary\_with(x1,x3,Null)$

P15: $\forall t,x1,x2\ PhiS\_of(x1,t) \wedge Nb\_of(x2,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant)$

P16: $\forall t,x1,x2,x3\ K1\_of(x1,t) \wedge Nb\_of(x2,t) \wedge Cox\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Contra\_variant)$

P17: $\forall t,x1,x2,x3,x4,x5\ K2\_of(x1,t) \wedge Cox\_of(x2,t) \wedge Nb\_of(x3,t)$
$\wedge Dj\_of(x4,t) \wedge Leff\_of(x5,t)$
$\rightarrow Vary\_with(x1,x2,Null) \vee Vary\_with(x1,x3,Null)$
$\vee Vary\_with(x1,x4,Null) \vee Vary\_with(x1,x5,Null)$

P18: $\forall t,x1,x2\ Eta\_of(x1,t) \wedge Nb\_of(x2,t)$
$\rightarrow Vary\_with(x1,x2,Contra\_variant)$

Figure 6.5.3 , continued

P19: $\forall t,x1,x2,x3\ Gmb\_of(x1,t) \wedge Gm\_of(x2,t) \wedge Gama\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$

P20: $\forall t,x1,x2,x3\ Gama\_of(x1,t) \wedge Nb\_of(x2,t) \wedge Cox\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \wedge Vary\_with(x1,x3,Contra\_variant)$

P21: $\forall t,x1,x2\ Ro\_of(x1,t) \wedge Lambda\_of(x2,t)$
$\rightarrow Vary\_with(x1,x2,Contra\_variant)$

P22: $\forall t,x1,x2,x3\ Lambda\_of(x1,t) \wedge Xd\_of(x2,t) \wedge Leff\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Contra\_variant)$

P23: $\forall t,x1,x2\ Xd\_of(x1,t) \wedge Nb\_of(x2,t)$
$\rightarrow Vary\_with(x1,x2,Contra\_variant)$

P24: $\forall t,x1,x2,x3\ Cgd\_of(x1,t) \wedge Cgd0\_of(x2,t) \wedge W\_eff(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$

P25: $\forall t,x1,x2,x3\ Cgd0\_of(x1,t) \wedge Cox\_of(x2,t) \wedge DL\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$

P26: $\forall t,x1,x2,x3,x4,x5\ Cgs\_of(x1,t) \wedge Cgs0\_of(x2,t) \wedge Cox\_of(x3,t)$
$\wedge W\_eff(x4,t) \wedge L\_eff(x5,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \wedge Vary\_with(x1,x3,Co\_variant)$
$\wedge Vary\_with(x1,x4,Co\_variant) \wedge Vary\_with(x1,x5,Co\_variant)$

P27: $\forall t,x1,x2,x3\ Cgs0\_of(x1,t) \wedge Cox\_of(x2,t) \wedge DL\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$

P28: $\forall t,x1,x2,x3\ Cgb\_of(x1,t) \wedge Cfox\_of(x2,t) \wedge DW\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$

P29: $\forall t,x1,x2\ Cfox\_of(x1,t) \wedge Tox\_of(x2,t)$
$\rightarrow Vary\_with(x1,x2,Contra\_variant)$

P30: $\forall t,,x1,x2,x3\ Csb(x1,t) \wedge Csb0(x2,t) \wedge Psi0\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$

P31: $\forall t,x1,x2,x3,x4\ Csb0\_of(x1,t) \wedge Ns\_of(x2,t) \wedge Nb\_of(x3,t)$
$\wedge Psi0\_of(x4,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$
$\vee Vary\_with(x1,x4,Contra\_variant)$

Figure 6.5.3 , continued

P32: $\forall\, t,x1,x2,x3\ Psi0\_of(x1,t) \wedge Ns\_of(x2,t) \wedge Nb\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$

P33: $\forall\, t,x1,x2,x3\ Cdb\_of(x1,t) \wedge Cdb0\_of(x2,t) \wedge Psi0\_of(x3,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$

P34: $\forall\, t,x1,x2,x3,x4\ Cdb0\_of(x1,t) \wedge Ns\_of(x2,t) \wedge Nb\_of(x3,t)$
$\wedge Psi0\_of(x4,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$
$\vee Vary\_with(x1,x4,Contra\_variant)$

P35: $\forall\, t,x1,x2,x3,x4\ Rsh\_of(x1,t) \wedge Mu\_of(x2,t) \wedge Ns\_of(x3,t)$
$\wedge Dj\_of(x4,t)$
$\rightarrow Vary\_with(x1,x2,Co\_variant) \vee Vary\_with(x1,x3,Co\_variant)$
$\vee Vary\_with(x1,x4,Co\_variant)$

Figure 6.5.3, continued

6.5.5. Making Inferences From the Mini Knowledge Base

The knowledge coded in Figure 6.5.3 could be the core of a knowledge base. By adding some facts about inference or measurement, it is possible for the system to derive other useful information. The following is an example to show how to make inferences from the knowledge base.

First, let's list some useful properties that would help the inference process. The Vary_with() relation is a transitive relation, meaning:

$$\textbf{A1: } \forall x1,x2,x3,q1,q2\ Vary\_with(x1,x2,q1) \wedge Vary\_with(x2,x3,q2) \wedge (q1 = Null \vee q2 = Null) \rightarrow Vary\_with(x1,x3,Null)$$

$$\textbf{A2: } \forall x1,x2,x3,q1,q2\ Vary\_with(x1,x2,q1) \wedge Vary\_with(x2,x3,q2) \wedge (q1 \neq q2) \wedge (q1 \neq Null) \wedge (q2 \neq Null) \rightarrow Vary\_with(x1,x3,Contra\_variant)$$

$$\textbf{A3: } \forall x1,x2,x3,q\ Vary\_with(x1,x2,q) \wedge Vary\_with(x2,x3,q) \wedge (q \neq Null) \rightarrow Vary\_with(x1,x3,Co\_variant)$$

From logic axioms we have:

If

$$p_1 \wedge p_2 \wedge \ldots \wedge p_i \rightarrow q_1 \vee q_2 \vee \ldots q_j \vee r$$

and

$$r \wedge t_1 \wedge \ldots \wedge t_m \rightarrow u_1 \vee u_2 \vee \ldots \vee u_n$$

Then

$$p_1 \wedge \ldots \wedge p_i \wedge t_1 \wedge \ldots \wedge t_m \rightarrow q_1 \vee \ldots \vee q_j \vee u_1 \vee \ldots \vee u_n$$

Where $r,p_1 \ldots p_i,q_1 \ldots q_j,t_1 \ldots t_m,u_1 \ldots u_n$ are all logic sentences. This inference rule listed above is defined as the *rule of resolution* [1].

Suppose a wafer contains an array of CMOS devices D1,D2...Dk. The PMOS (or NMOS) transistor part of any CMOS device d is represented by d_P (or d_N). In addition, the parameter α of any transistor t is represented by t_α. For example, D1_N_Gm denotes the transconductance (g_m) of transistor D1_N, which is the NMOS part of CMOS device D1.

With the representation method above, we can include facts about the wafers into our knowledge. The set of rules is called the wafer specification, and they are of the following forms:

S1: $\rightarrow Cmos(D1)$
S2: $\rightarrow Pmos\_of(D1\_P,D1)$
S3: $\rightarrow Nmos\_of(D1\_N,D1)$
S4: $\rightarrow Gm\_of(D1\_P\_Gm,D1\_P)$
.
.

With rules P1 to P35, A1 to A3, and wafer specification S1,S2..., we can deduce other facts not listed in the knowledge base by using the resolution method. For example, one can deduce the fact:

$\rightarrow Vary\_with(D1\_N\_Gm,D1\_N\_Mu)$

from this knowledge base by using only the resolution method.

This knowledge base can also be used to specify which measurements should be made. If we define relations Check(α), which means executing the checking routine for element α; Normal(α), which means element α is normal; and Abnormal(α), which means element α is abnormal, then we can add a rule to conduct measurements:

$\forall x1,x2,q\ Abnormal(x1) \wedge Vary\_with(x1,x2,q)$
$\rightarrow Check(x2)$

The result of Check(x2) might be either Abnormal(x2) or Normal(x2), which could be added to the knowledge base as new facts.

The above examples show only two applications of the knowledge represented in Figure 6.5.3. There are many other possible applications, which we don't intend to describe here.

6.5.6. Future Work

We described above the ontology for CMOS SPICE parameters and presented basic knowledge for the relations among parameters. As mentioned previously, this ontology study is paving a road for future activities, which will include a similar analysis of the remaining relevant IC processing parameters, and an assortment of test structure required to provide a complete set of parametric data measurary for correct problem diagnosis using the alluded to expert systems.

References

[1] Michael Genesereth and Nils Nilsson.
Fundamentals of Artificial Intelligence.
, 1986.

[2] Paul R. Gray and Robert G. Meyer.
Models for Integrated-Circuit Active Devices.
Analysis and Design of Analog Integrated Circuits.
John Willey & Sons, 1986, Chapter 1.

[3] David A. Hodges and Horace G. Jackson.
Analysis and Disign of Digital Integrated Circuits.
McGraw-Hill, Inc., 1983.

[4] John Mohammed and Reid Simmons.
Qualitative Simulation of Semiconductor Fabrication.
March, 1986.

[5] Jeff Yung-Choa Pan and Jay M. Tenenbaum.
P.I.E.S.: An Engineer''s Do-It-Yourself Knowledge System for Interpretation of parametric Test Data.
AI Magazine 7(4):62-69, 1986.

[6] K. N. Ratnakumar and James D. Meindl.
Short-Channel MOST Threshold Voltage Model.
IEEE Journal of Solid-State Circuit SC-17(5):927-948, October, 1982.

[7] B. J. Sheu et al.
Berkeley Short-Channel IGFET Model (BSIM).
Technical Report, Electronics Research Laboratory, U.C.Berkeley, March, 1984.

[8] T. Quarles et al.
SPICE 3A7 User''s Guide
1986.

END

DATE

3-88

DTIC